Culc- H05961-5-KP665

THE JOURNAL OF
PHILOSOPHY

Calcutta University Library
ALIPURE
CAMPUS

VOLUME XCIV, NUMBER 1
JANUARY–DECEMBER 1997

CALCUTTA UNIVERSITY LIBRARY

Edited by
Bernard Berofsky, Akeel Bilgrami
Arthur C. Danto, Kent Greenawalt
Isaac Levi, Mary Mothersill

⑤

Published by The Journal of Philosophy, Inc.

KP-665
(12.4.2001).

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. In 1923, the JOURNAL was incorporated in the State of New York under its present name. Since its founding, it has been published from Columbia University and printed by the Lancaster Press in Lancaster, Pennsylvania.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 1
JANUARY 1997

page
5 *Infinite Value and Finitely Additive Value Theory* Peter Vallentyne and Shelly Kagan

REVIEW ESSAYS

27 Gregory Vlastos: *Socratic Studies* Martha C. Nussbaum

BOOK REVIEWS

46 Axel Honneth: *The Struggle for Recognition: The Moral Grammar of Social Conflicts* Michael O. Hardimon

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER. Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

Calcutta University Library
ALIPORE CAMPUS

VOLUME XCIV, NUMBER 2
FEBRUARY 1997

CALCUTTA UNIV. LIBRARY

page

55 *Analog Representation beyond Mental Imagery* James Blachowicz

COMMENTS AND CRITICISM

85 *Elgin on Lewis's Putnam's Paradox* Bas C. van Fraassen

BOOK REVIEWS

94 Quentin Skinner: *Reason and Rhetoric in the Philosophy of Hobbes* David Gauthier
97 Daniel C. Dennett: *Kinds of Minds: Toward an Understanding of Consciousness* Nicholas Humphrey
103 William P. Alston: *A Realist Conception of Truth* Timothy Williamson
107 NEW BOOKS: ANTHOLOGIES
108 NEW BOOKS: TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIV LIBRARY

VOLUME XCIV, NUMBER 3
MARCH 1997

page
109 *The Rule of Adjunction and Reasonable Inference* Henry E. Kyburg, Jr.

126 *Fallibilism, Ambivalence, and Belief* Jonathan Roorda

BOOK REVIEWS
156 John Martin Fischer: *The Metaphysics of Free Will: An Essay on Control* R. Jay Wallace
160 Elliott Sober: *From a Biological Point of View: Essays in Evolutionary Philosophy* Peter Godfrey-Smith

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

CALCUTTA UNIV LIBRARY

VOLUME XCIV, NUMBER 4
APRIL 1997

page
165 *The Epistemic Status of Probabilistic Proof* Don Fallis
187 *Fainthearted Conditionals* Michael Morreau

BOOK REVIEWS
212 Bernard Berofsky: *Liberation from Self: A Theory of Personal Autonomy* Gerald Dworkin

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY

VOLUME XCIV, NUMBER 5
MAY 1997

page
217 *Probability as a Guide to Life* Helen Beebee and David Papineau
244 *A Shooting-room View of Doomsday* William Eckhardt

COMMENTS AND CRITICISM
260 *Dion's Foot* Eric T. Olson

BOOK REVIEWS
266 Stephen Darwall: *The British Moralists and the Internal 'Ought': 1640-1740* Nicholas L. Sturgeon

272 NEW BOOKS: TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus:* Sidney Morgenbesser. *Consulting Editors:* David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor:* Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President,* Arthur C. Danto; *Vice President,* Akeel Bilgrami; *Secretary,* Leigh S. Cauman; *Treasurer,* Barbara Gimbel; *Other Trustees:* Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY

VOLUME XCIV, NUMBER 6
JUNE 1997

page
273 *The Right and the Good* Judith Jarvis Thomson
299 *Temporal Horizons of Justice* Bruce Ackerman

BOOK REVIEWS
318 Béatrice Longuenesse: *Kant et le pouvoir de juger: Sensibilité et discursivité dans l'analytique transcendentale de la* Critique de la raison pure. Robert B. Pippin

325 NEW BOOKS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 7
JULY 1997

page
329 *The Mathematical Structure of the World: The World as Graph* Randall R. Dipert

REVIEW ESSAYS
359 Pasquale Frascolla: *Wittgenstein's Philosophy of Mathematics*; and Christoffer Gefwert: *Wittgenstein on Philosophy and Mathematics* Michael Dummett

374 NOTES AND NEWS
375 NEW BOOKS
378 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 8
AUGUST 1997

page
381 *When a Pain Is Not* — Valerie Gray Hardcastle
410 *Reflective Knowledge in the Best Circles* — Ernest Sosa

BOOK REVIEWS
431 Martha C. Nussbaum: *Poetic Justice: The Literary Imagination and Public Life* — Richard Eldridge

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 9
SEPTEMBER 1997

page
435 *Lying, Deceiving, or Falsely Implicating* — Jonathan E. Adler
453 *Putnam's Born-again Realism* — Fred Sommers
472 *Thick Epistemic Access: Distinguishing the Mathematical from the Empirical* — Jody Azzouni

BOOK REVIEWS
485 Charles Larmore: *The Morals of Modernity* and *The Romantic Legacy* — Alasdair MacIntyre

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 10
OCTOBER 1997

page
491 *Liberalism, Sanctity, and the Prohibition of Abortion* John Tomasi
514 *Vague Identity and Fuzzy Logic* B. Jack Copeland
535 *The Lot of the Causal Theory of Mental Content* Robert Cummins

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY
CAMPUS

VOLUME XCIV, NUMBER 11
NOVEMBER 1997

page
543 *Quine's Limited Naturalism* Robert J. Fogelin
564 *Manifest Kinds* Mark Johnston

BOOK REVIEWS
584 Kent Greenawalt: *Private Consciences and Public Reasons* Robert Amdur
588 Frederic Schick: *Making Choices: A Recasting of Decision Theory* Isaac Levi

597 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NUMBER 12
DECEMBER 1997

page
599 *Why Constitution Is Not Identity* Lynne Rudder Baker
622 *How to Keep the 'Physical' in Physicalism* Andrew Melnyk

BOOK REVIEWS
638 Stephen L. White: *The Unity of the Self* Brian P. McLaughlin

645 INDEX TO VOLUME XCIV

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the editors and trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1997

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; printed by Lancaster Press, Inc., Lancaster, PA. Typeset by Tapsco, Inc., Akron, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1997 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

CALCUTTA UNIVERSITY LIBRARY

THE JOURNAL OF PHILOSOPHY

CALCUTTA UNIVERSITY LIBRARY

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 1997 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices. Send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, NY, NY 10027.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NO. 1, JANUARY 1997

INFINITE VALUE AND FINITELY ADDITIVE VALUE THEORY*

Call a theory of the good—be it moral or prudential—*aggregative* just in case (1) it recognizes local (or location-relative) goodness, and (2) the goodness of states of affairs is based on some aggregation of local goodness. The locations for local goodness might be points or regions in time, space, or space-time; or they might be people or states of nature.[1] Any method of aggregation is allowed: totaling, averaging, measuring the equality of the distribution, measuring the minimum, and so on. Call a theory of the good *finitely additive* just in case it is aggregative, and for any finite set of locations it aggregates by adding together the goodness at those locations. Standard versions of total utilitarianism typically invoke finitely additive value theories (with people as locations).

A puzzle can arise when finitely additive value theories are applied to cases involving an infinite number of locations (people, times, and so on). Suppose, for example, that temporal locations are the locus of value, and that time is discrete and has no beginning or end.[2]

---

* Thanks to Dick Arneson, Tim Bays, John Broome, James Cain, Tyler Cowen, Chuck Cross, Michael Della Rocca, Jorge Garcia, Brad Hooker, Jerome Keisler, Mark Nelson, Alastair Norcross, Teddy Seidenfeld, Ted Sider, Howard Sobel, Roy Sorensen, and Paul Weirich for helpful discussion and comments.

[1] Here we follow John Broome's usage of the term 'location' as the generic term for the things with which local goodness is associated; see his *Weighing Goods* (Cambridge: Blackwell, 1991).

[2] Here and below, we shall assume for our examples that (1) there are only denumerably many locations in space, time, and space-time, and (2) they are discrete in the sense that for any location and any direction there is a well-defined "next" location. These assumptions are made for simplicity of presentation, and play no role in our argument.

0022-362X/97/9401/5–26

How would a finitely additive theory (for example, a temporal version of total utilitarianism) judge the following two worlds?

Goodness at Locations (for example, times)
w1: ..., 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
w2: ..., 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
Example 1

At each time, w1 contains 2 units of goodness and w2 contains only 1. Intuitively, we claim, if the locations are the same in each world, finitely additive theorists will want to claim that w1 is better than w2. But it is not clear how they could coherently hold this view. For using standard mathematics, the sum of each is the same infinity, and so there seems to be no basis for claiming that one is better than the other.[3] (Appealing to Cantorian infinities is of no help here, since for any Cantorian infinite N, $2 \times N = 1 \times N$.)

We shall argue that such theories can and should judge some worlds with an infinite number of locations as better than others. Total utilitarianism, for example, can and should judge w1 as better than w2 when they have the same locations. Moreover, we shall argue that there are some perfectly general metaprinciples governing how finitely additive theories should make judgments when an infinite number of locations are involved.[4]

---

[3] This problem does not always arise when there are an infinite number of locations. First, if the locations are ordered (as in time or space) and bounded (entirely inside some finite region), then there may be no problem. For example, the "sum" (integral) of the finite values at all the infinitely many real-numbered locations in some bounded interval (for example, 1 to 2) is well defined, and finite. Second, if the locations are unbounded, but their values asymptotically approach zero "sufficiently quickly" (for example, as in $v(x) = 1/x^2$), then the "sum" (integral) may be well defined and finite. In these cases, the standard integration techniques deal "almost perfectly" with aggregation over an infinite number of locations. Strictly speaking, it is not perfect because standard integration holds that the aggregate value of 1 unit of happiness at each point in time inclusively between two points in time (with time being real valued) is the same as that of 1 unit at each time in this interval except with 2 units of happiness at the endpoint. Standard integration assigns the same aggregate value to these functions, but the second, we would argue, has a greater value. We agree, of course, that the results of standard integration are infinitesimally close to the correct answer, and for the purposes of this paper, we shall ignore the small errors. Thus, all references to sums should be understood as including integration when this is appropriate.

[4] The problem of aggregating when time is infinite in length has been discovered and given roughly the same solution (on which we generalize in the present paper) at least three times. Most recently, it was discovered in Mark Nelson, "Utilitarian Eschatology," *American Philosophical Quarterly*, XXVIII (1991): 339-47; and addressed in Vallentyne, "Utilitarianism and Infinite Utility," *Australasian Journal of Philosophy*, LXXI (1993): 212-17. Prior to that, it was discovered and addressed in (the important but unnoticed!) Krister Segerberg, "A Neglected Family of Aggre-

We shall assume throughout that the amount of goodness at each location is finite (so that the infinite goodness comes from the infinite number of locations and not from the values at the locations). We shall further assume that goodness is fully interlocationally comparable in the sense that it is possible to ensure that the same scale (both the zero point and the unit) for measuring goodness is used for all locations. This is a presupposition of finitely additive value theories. For if the numbers are not on the same scale, it makes no sense to add them together.[5] All goodness numbers are thus assumed to be on the same scale.

Finally, because for generality we make no specific assumption about what the relevant basic locations of goodness are (people, times, and so on), all references to locations should be understood as those which are specified by a given finitely additive theory under consideration.[6]

I. NONSTANDARD NUMBERS

Before proceeding, we need to mention a possible partial solution to the puzzle. In the discussion of the example above, we assumed that the sum of a denumerably infinite number of 2s is the same as the sum of an equal number of 1s. On standard mathematics, it would be more correct to say that neither sum is well defined, but the point would remain that neither sum would be defined and greater than the other. The important point to note here is that there is such a thing as nonstandard mathematics (for example, nonstandard analy-

---

gation Problems in Ethics," *Noûs*, x (1976): 221-44. And prior to that, it was discovered and addressed by Frank Ramsey in the 1920s and addressed by various economists, especially since the 1960s; see Carl Christian Von Weizsacker, "Existence of Optimal Programs of Accumulation for an Infinite Time Horizon," *Review of Economic Studies*, XXXII (1965): 85-104; and David Gale, "On Optimal Development in a Multi-Sector Economy," *Review of Economic Studies*, XXXIV (1967): 1-18. See also the discussions (and many other references) in D.E. Campbell, "Impossibility Theorems and Infinite Horizon Planning," *Social Choice and Welfare*, II (1985): 283-93; L.G. Epstein, "Intergenerational Preference Orderings," *Social Choice and Welfare*, III (1986): 151-61; and Luc Lauwers, *Social Choice with Infinite Populations*, Dissertation 101 (Catholic University of Leuven (Monitoraat E.T.E.W, Dekenstraat 2, B-3000 Leuven, Belgium), 1995).

[5] If the locations (for example, people) are the same, then only unit comparability is needed, since the zeros will cancel out. But in the more general case, where the locations may be different in the worlds being evaluated, zero comparability is also needed.

[6] For example, if, according to a given finitely additive theory of value, people are the only ultimate bearers of value, then only personal locations (and not temporal ones) are referenced. For a discussion of the problems that arise if locations are not restricted to ultimate locations, see James Cain, "Infinite Utility," *Australasian Journal of Philosophy*, LXXIII (1995): 401-04; and Vallentyne, "Utilitarianism and Infinite Utility," *Australasian Journal of Philosophy*, LXXI (1993): 212-17.

sis) dealing with infinitesimals and infinite numbers. Although such mathematics is not mainstream, it has been recognized as legitimate since the work of Abraham Robinson in the 1960s established that the existence of such numbers is perfectly consistent with standard mathematics.[7] We shall now briefly explain how appealing to such mathematics may offer a solution to the puzzle.

Standard mathematics does not recognize infinitesimals and infinite numbers for the purposes of normal addition or multiplication. Nonstandard mathematics, however, does. Indeed, on nonstandard mathematics, all the basic arithmetic operations are well defined and operate in the "usual way" for both finite and infinite numbers. It recognizes, for example, that adding or multiplying two positive infinite numbers always yields a result that is greater than either of the two original numbers. It also says that, if N is some infinite integer and N 1s are added together, then the result is N. If one more 1 is added to the total, the result is $N+1$, which is 1 greater than N. If N 2s are added together, the result is $2\times N$, which is twice as great as N. N, $N+1$, and $2\times N$ are each nonstandard infinite numbers, and the first is smaller than the second, which is smaller than the third.

If the infinities involved in the puzzle cases are bounded nonstandard infinities, then nonstandard mathematics provides a straightforward resolution of the puzzle of the above example. For although the two sums each involve the same nonstandard infinite number of locations (call it N), the first nonstandard sum (of w1) is greater than the second (of w2), since $2\times N > N$.

Is this the end of the story? No. For nonstandard mathematics, like standard mathematics, is silent in many cases where there are *unboundedly* many (that is, more than any nonstandard infinite number) positive numbers that are summed. Nonstandard mathematics says, for example, that 2 units of goodness at (nonstandard infinite) H locations (that is, <2,2,2,...2>, where there are H 2s) has a greater total than 1 at H locations (that is, < 1,1,1...1>, where there are H 1s). But it, like standard mathematics, is silent about the comparisons of 2, versus 1, at each of unboundedly many locations (that is, <2,2,2,...> versus <1,1,1,...>).

So, nonstandard mathematics resolves the puzzle where the number of locations is a (bounded) nonstandard number, but not in general when it is unbounded. We shall therefore focus on the unbounded case, and for simplicity we shall assume that only standard numbers are involved. We shall develop and defend a metaprinciple for aggregating the goodness of infinitely many locations which makes no essential use of nonstandard mathematics.

---

[7] See Abraham Robinson, *Non-Standard Analysis* (Amsterdam: North Holland, 1966). For an extremely intuitive and accessible introduction to nonstandard mathematics, see H.J. Keisler, *Elementary Calculus* (Boston: Prindle, Weber, and Schmidt, 1976), chapter 1.

II. THE BASIC IDEA: AGGREGATING WITHOUT APPEALING TO ANY ESSENTIAL NATURAL ORDER OF LOCATIONS

We start by formulating a metaprinciple that is applicable to all sorts of locations—whether or not they have any essential natural order (a notion that is explained below). In the following sections, we strengthen this metaprinciple for cases where there is some essential natural order (as arguably there is for spatial and temporal locations but not for people and states of nature).

Our most basic idea is very simple, and can be illustrated using example 1—in which w1 has 2 units of goodness at each location and w2 has 1 unit at each. Suppose that these two worlds contain exactly the same locations (for example, the same people). Finitely additive theories should, we claim, judge w1 as better than w2 for the following reason: no matter what finite set of locations one considers, relative to that set w1 has a greater total than w2. That is, our most basic idea (BI) is:

> BI (basic idea): if w1 and w2 have exactly the same locations, and if, relative to any finite set of locations, w1 is better than w2, then w1 is better than w2.[8]

This is the core of our approach, and we hope it is sufficiently plausible to need little discussion at this point. (We shall, of course, return to it below.) As it stands, it is incredibly weak: it has implications for finitely additive theories only where (1) the locations in the two worlds are exactly the same and (2) one world has more goodness at every single location (since just one countervailing singleton set renders the idea silent). It is hardly a robust metaprinciple, but it is fine, we claim, as far as it goes.

The rest of the paper is concerned with strengthening BI. There are two sorts of strengthening that we shall develop: one extends BI by weakening the requirement that one world be better, relative to all finite sets of locations; the other strengthens the metaprinciple to cover cases where the locations are not exactly the same.

---

[8] For simplicity, we shall assume throughout that for the purposes of aggregation (1) all locations of a given type (for example, time, or spatial locations, or persons) have the same weight (both within and across worlds), and (2) the total weight of the locations in a given volume (or area, or interval, depending on the case) is—where defined—also constant within and across worlds. If these assumptions are dropped, then an appropriate condition needs to be added to our metaprinciples.

Consider the following example of two worlds having the same locations:

Goodness at Locations (for example, times)
w1: ..., 2, 2, 2, 2, 1, 2, 2, 2, 2, ...
w2: ..., 1, 1, 1, 1, 2, 1, 1, 1, 1, ...
Example 2

Here and throughout, when the locations are the same in the two worlds, we display them in the same column.

BI is silent here, since there is a finite set of locations, namely, the singleton set with the location that has 1 in w1 and 2 in w2, relative to which w2 is better than w1 (and, of course, there are sets relative to which w1 is better than w2). Are there no grounds for finitely additive theories to judge w1 as better than w2? Surely there are! For any finite set of locations containing at least three locations will be such that w1 is better than w2 relative to that set. And for cases such as the one above, that is surely sufficient to judge w1 as better.

The point here is that it is not necessary for judgments of betterness that one world be better than a second relative to all finite sets of locations. It is sufficient for this sort of case, it seems, that any finite set can be expanded sufficiently so that, relative to all further finite expansions, the first world is better. The rough idea is that, if, no matter what finite set of locations you start with, you can expand enough so that, relative to all further expansions, one world is better than another, then the former world is better tout court than the latter world. (In example 2, no matter what finite set one considers, if one expands it to include at least two locations, w1 is better than w2, relative to all further expansions.)

One strengthening of BI, which is very close to being plausible, is:

> RSBI (rejected strengthened basic idea): if (1) w1 and w2 have exactly the same locations, and (2) for any finite set of locations there is a finite expansion (superset) such that, relative to all further finite expansions w1 is better than w2, then w1 is better than w2.

As indicated above, RSBI rightly judges w1 as better than w2 in example 2. Indeed, for almost all cases, RSBI gives clearly plausible directives. Unfortunately, there is one sort of technical case where RSBI's directives are incorrect. Consider the following example:

Goodness at Locations (for example, times)
w1: 1, 0, 0, ...
w2: 1/2, 1/4, 1/8, ...
Example 3

Here, assuming that the locations are the same in the two worlds, any finite set of locations can be expanded by adding the 1-over-1/2 location (if it is not already included), and, relative to any further finite expansion, w1 is better than w2. Thus, RSBI directs that w1 be judged as better. But according to both standard and nonstandard mathematics, the total of 1/2+1/4+1/8... is 1, and so the two worlds contain the same amount of goodness.

The problem that arises for RSBI in example 3 is that, although no matter how one finitely expands a set containing the 1-over-1/2 location, w1 is better than w2 relative to that set, the relative advantage of w1 over w2 decreases the more one expands the set. At the limit, when the set is expanded to include all (infinitely many) locations, the advantage disappears. To deal with this case, let us say that, relative to a finite set of locations, a world, w1, is *k-better* than a second world, w2, for some positive number *k*, just in case the overall goodness of these locations in w1 is at least *k* units better than their overall goodness in w2. (Here, for simplicity, we leave implicit a reference to some fixed scale of goodness.)

We can now formulate a strengthening of BI that we endorse:

> SBI1 (strengthened basic idea 1): if (1) w1 and w2 have exactly the same locations, and (2) for any finite set of locations there is a finite expansion and some positive number, *k*, such that, relative to all further finite expansions, w1 is *k*-better than w2, then w1 is better than w2.

SBI1 is silent in example 3. For no matter what number, *k*, is chosen (no matter how small), and no matter how an initial finite set is expanded (even if the 1-over-1/2 location is added), one can further expand by adding enough of the remaining locations with the largest values in w2 so that the total in the w1 locations is *not* at least *k* units of goodness greater than that in w2 (and vice versa). And in example 2, SBI1 correctly directs finitely additive theories to judge w1 as better than w2 (since for any finite set containing at least three locations w1 will have a total that is at least 1-better than w2).

In what follows, we shall use the '*k*-better' terminology in the official statements of our metaprinciples, but for stylistic reasons we shall leave it to be understood implicitly in our discussions. Readers unconcerned with cases where betterness disappears at the limit may safely ignore *k*-betterness and think solely in terms of betterness.

For locations (like people and states of nature, but perhaps unlike spatial and temporal locations) that have no essential natural order (a notion that we shall explain below), SBI1 says all that we have to say. We shall now strengthen SBI1 for cases where locations have certain sorts of essential natural order.

III. AGGREGATING WHEN LOCATIONS HAVE AN ESSENTIAL NATURAL ORDER AND ARE THE SAME

SBI1 has bite only when every finite set of locations can be expanded enough so that, relative to all further finite expansions, one world is better than another. We think that SBI1 can be plausibly strengthened by making it applicable to a wider range of cases. Consider the following example:

Goodness at Locations (for example, times)
w1: ..., 5, 1, 5, 1, 5, 1, 5, 1, 5, ...
w2: ..., 3, 2, 3, 2, 3, 2, 3, 2, 3, ...
Example 4

In this example, no matter what finite set one selects, and no matter how one finitely expands initially, there are further finite expansions relative to which w1 is better (one just expands by adding enough 5-over-3 locations), and there are finite expansions relative to which w2 is better (one just expands by adding enough 1-over-2 locations). Thus, SBI1 is silent.

Nonetheless, it may be appropriate for finitely additive theories to judge w1 as better than w2. For if the (for example, temporal) locations have an essential natural order (as characterized below), then not all expansions of a given set are normatively relevant. For example, if one started with a singleton 5-over-3 set, then no expansion that includes another 5-over-3 location without including the intermediate 1-over-2 locations is normatively relevant. For such expansions violate the essential natural order of the locations (they skip over locations). Thus, we claim, if the locations have an essential natural order, as arguably spatial and temporal locations do, then we can strengthen our metaprinciple by considering only certain kinds of order-respecting expansions.

In order to develop this idea, we need to explain the idea of locations having an order and the idea of this order being both natural and essential. We postpone discussing the notions of naturalness and essentialness until after it is clear how they will be used, and start by explaining the idea of locational order.

The notion of locational order that we have in mind is that of a topological manifold. We shall not define it precisely, but the rough idea is that locations are connected to each other so that the notion of a (continuous, or unbroken) path is well defined and all locations are path connected. Loose marbles, for example, have no order, but they do when glued in a line on a rubber yardstick so that every marble touches its two adjacent neighbors. Movement from the marble at one end of the yardstick to the marble at the other end via adjacent marbles traces a path, but movement from one marble to a nonadjacent one is not a path (since there is a gap). This idea of locational order does not require there to be a well-defined, or fixed, distance between any two locations, since the paths (as on a rubber yardstick) may be contractible, expandable, or twistable.

If there is locational order (topological structure)—as with spatial and temporal locations, for example—the notion of a ***bounded region*** is well defined. Roughly, a bounded region is a set of locations that are all "inside a boundary." For the one-dimensional case, a bounded region is a set of all the locations inclusively between some two locations (for example, the interval from 2 to 4 on the real line). For the two-dimensional case, a bounded region is a set of all the locations inclusively inside some simple closed path (for example, circle or rectangle). (A simple closed path (curve) is a path starting and ending at the same point without passing through any other locations twice.)

Locations with such topological order may either be ***discretely connected*** (that is, such that there are well-defined adjacent locations, and thus only finitely many locations between two locations on a given path), or ***densely connected*** (that is, infinitely many locations between any two locations on a given path).[9] A bounded region contains finitely many locations in the discrete case but infinitely many in the dense case.

For expositional brevity, we are going to engage in some double talk. For from now we shall use the term 'bounded region' in a looser sense than that just defined. We shall understand a ***bounded region*** in this loose sense to be (1) a bounded set in the strict sense, if there is an essential natural locational order, and (2) any finite set, if there is no essential natural order. This will permit us to use the same language to cover both sorts of cases.

We claim that, where there is an essential natural order, not all logically possible finite expansions need to be considered. Only ***bounded regional expansions*** (that is, expansions that are bounded regions in the strict sense as defined by the topology) need be considered. SBI1 can, we claim, be strengthened as follows:

---

[9] For the discrete case, we assume that each location has only a finite set of adjacent locations, and any simple closed path (that is, sequence of adjacent locations with no repetition except that the first location is also the last; for example, a circle or square) separates the set of all other locations into disconnected sets (that is, such that no location inside is adjacent to any location outside). This condition holds automatically in the case of dense topology.

> SBI2 (strengthened basic idea 2): if (1) w1 and w2 have exactly the same locations, and (2) for any bounded region of locations there is a bounded regional expansion and some positive number, *k*, such that, relative to all further bounded regional expansions w1 is *k*-better than w2, then w1 is better than w2.[10]

Applied to example 4, this metaprinciple is silent, if the locations have no essential natural order (for example, if they are people), but it directs finitely additive theories to judge w1 as better if the locations are the same, have an essential natural order (for example, if they are temporal), and are listed in their natural order. For no matter what bounded region one starts with, if one expands finitely leftward or rightward (or both) with no gaps (a bounded regional expansion) so as to include at least two 5-over-3 locations, then relative to any further such expansion w1 is better than w2. The fact that it is logically possible to expand a given finite set of locations by adding *only* 1-over-2 locations (and thus produce a higher total for w2) is, we claim, irrelevant, because this violates the essential natural order of temporal locations (by skipping over locations). SBI2 rightly directs that w1 be judged better than w2.[11]

Below, we shall further strengthen this metaprinciple. First, however, we need to say something about a location order being essential and natural.

Naturalness is the most difficult notion—indeed, we are embarrassed to say that we cannot give a crisp definition of what we mean by it! But the intuitive idea, we think, can be made reasonably clear. Locations can be ordered in all sorts of ways. People, for example, can be ordered by their dates of birth, their dates of death, their first names, their last names, and so on. At most one of these orderings is natural in the sense of reflecting an ontologically fixed order. In the case of people, we are inclined to think there is no natural order at all. Temporal and spatial locations, on the other hand, may well have natural orderings in this sense. For at least on a common conception of time and space, it seems plausible to think that they are ontologically fixed in a certain order. You cannot "get" from year 1994 to year 2000 without "passing through" the intermediate years.

---

[10] SBI2 is a generalization of a principle (PMU*) formulated and defended in Vallentyne, "Utilitarianism and Infinite Utility." It is a generalization in three respects: first, it applies to any finitely additive theory of the good, and not just utilitarianism; second, it is applicable to any sort of location that has an essential natural order—not just time; third, it is applicable to worlds that extend infinitely in all directions, and not just to those which are infinite only "toward the future."

[11] When the locations are discretely connected, SBI2 can have implications without there being a distance metric. This is because a bounded region will have only finitely many locations, and their values can simply be added together. When the locations are densely connected, however, SBI2 has no implications without a distance metric. For a distance metric is required to aggregate the infinitely many locations in a dense bounded region (without a distance metric, the dense interval of 1 to 2 is indistinguishable in "size" from the dense interval from 1 to 20, for example).

Strictly speaking, we want to allow the naturalness of an ordering to be world relative. But for simplicity we shall consider only natural orderings that are essential, that is, the same in all worlds. Although it seems somewhat strange to think of natural ordering as being nonessential, we do not see how to rule out this possibility. Instead, we shall simply restrict the scope of our metaprinciple below to essential natural orderings. At least where there is an essential natural order, it is, we shall argue, possible to strengthen the metaprinciple above.

We hope that these sketchy remarks on essential natural orderings are enough to indicate that there is something to this notion. We acknowledge that the notion remains somewhat mysterious, but shall not attempt here to make it more precise. Below, we shall typically assume that temporal and spatial locations do, and people do not, have an essential natural order; but this is for the sake of illustration only. Strictly speaking, our strengthenings of SBI1 apply to whatever locations have an essential natural order. If it turns out that no locations have such order, then the strengthenings have no applicability.

Although SBI2 is an improvement over SBI1, it can be strengthened further. To see this consider the following two-dimensional example (where there is one unit of goodness at each location except for two rows of -1):[12]

Goodness at Locations

$$\begin{array}{ccccccc}
 & \text{w1:} & \ldots\ldots\ldots\ldots & & & & \\
\ldots, & 1, & 1, & 1, & 1, & 1, & 1, \ldots \\
\ldots, & 1, & 1, & 1, & 1, & 1, & 1, \ldots \\
\ldots, & -1, & -1, & -1, & -1, & -1, & -1, \ldots \\
\ldots, & -1, & -1, & -1, & -1, & -1, & -1, \ldots \\
\ldots, & 1, & 1, & 1, & 1, & 1, & 1, \ldots \\
\ldots, & 1, & 1, & 1, & 1, & 1, & 1, \ldots \\
 & & \ldots\ldots\ldots\ldots & & & &
\end{array}$$

Example 5

Is w1 better than *its zero world* (that is, a world with the same locations but with 0 goodness at each location)? SBI2 is silent here, even assuming that the locations have an essential natural order (with each location being adjacent to just the locations immediately above, below, and to the side). For no matter what bounded region one starts with, and no matter how one finitely regionally expands initially, one can regionally expand along the two -1 rows enough to produce a negative total, and one can also regionally expand along two adjacent columns enough to produce a positive total. Consequently, SBI2 is silent. We think, however, that w1 is indeed better than its zero world. After all, any two rows of 1s will cancel out the two rows of -1 and there will be infinitely many rows of 1s left.

---

[12] The example is a modification of one provided by an anonymous editor of this JOURNAL, which helped us see the need for a strengthening.

SBI2 is silent because it considers all bounded regional expansions relevant—no matter how selective they are in the directions along which they expand. The expansion along the rows of -1s, for example, is deemed relevant (and is enough to block SBI2 from directing that w1 be judged better than w2). We claim that SBI2 can be plausibly strengthened by restricting its appeals only to what we call uniform expansions, which expand in all directions "uniformly."

If there is an essential natural distance metric for locations, then a uniform expansion is simply one that adds a band of constant width to the initial region. If there is no essential natural distance metric, but locations are discretely connected, then a uniform expansion can be defined as follows. For the one-dimensional discrete case a uniform expansion of a given region (interval) is simply an expansion that adds on the same number of locations to the right as to the left. For the two-dimensional discrete case a uniform expansion of a given region is a region obtainable from the initial one by successively adding on layers of locations, where a layer around a given region is a smallest (in terms of the number of locations) simple closed path that has the given region strictly on the inside. A *simple closed path* (for the discrete case) is a sequence of adjacent locations that starts and ends at the same location with no repetition of intermediate locations (for example, a circle or rectangle). There may be more than one smallest simple closed path with the initial region strictly on the inside, and in that case all such paths are layers.[13]

Where locations have no essential natural order, or where their essential natural order is dense but without an essential natural distance metric, there is no appropriate way of strengthening SBI2. For expositional brevity, in such cases we shall engage in some double talk, and understand a uniform expansion to be any finite superset.

---

[13] For simplicity, we assume that the locations of a world are unbounded in all directions. Given that the problem of infinite aggregations can arise whenever the locations are unbounded in some directions (even if not all), this is a simplification. If the assumption is dropped, the definition of a layer around a given region would need to be modified to read "a smallest simple closed path that has the initial region strictly on the inside, except perhaps for those locations in the initial region that are on a world boundary" (to recognize that no expansion is possible in certain directions beyond a boundary). Also recall that for simplicity we are assuming throughout that all locations in all worlds have the same weight, and that the weight of the locations per unit of volume/area/distance is uniform. If this assumption is dropped, the definition of layers would need to appeal to expansions of uniform weight in all directions.

We can now strengthen SBI2 to:

> SBI3 (strengthened basic idea 3): if (1) w1 and w2 have exactly the same locations, and (2) for all bounded regions of locations there is a bounded uniform expansion and a positive number, *k*, such that, relative to all further bounded uniform expansions, w1 is *k*-better than w2, then w1 is better than w2.[14]

This is the same as SBI2 except that the appeal to regional expansions has been replaced by an appeal to uniform expansions.

In example 5 above, SBI3 rightly directs finitely additive theories to judge w1 as better than its zero world (where all locations have zero goodness). For no matter what bounded regions one starts with, if one regionally expands uniformly until the region contains more 1s than −1s, then any further finite uniform expansion will contain more 1s than −1s (and thus be better than its zero world). Thus, SBI3 directs that finitely additive theories judge w1 as better than its zero world. This judgment, we claim, is plausible.[15]

---

[14] A slightly stronger principle is the following: if (1) two worlds have exactly the same locations, (2) there is a bounded region of locations, and a positive number, *k*, such that relative to all bounded uniform expansions, w1 is *k*-better than w2, but (3) not vice versa (that is, (2) with w1 and w2 permuted), then w1 is better than w2. This is like the principle in the text except that it requires only that there be *some* initial region relative to which all bounded uniform expansions favor w1, and in that it has a 'and not vice versa' clause (needed to avoid contradictory judgments). It directs finitely additive theories to judge <... 1,−1,1,−1,2,−1,1,−1,1...> as better than its zero world (since, relative to all bounded uniform expansions of the <−1,2> region, w1 is better than w2, and the not vice versa clause is satisfied). SBI3, on the other hand, is silent (since no bounded uniform expansion of a −1 location makes w1 better for all further bounded uniform expansions). (Both principles are silent if the 2 is replaced by a 1, and both judge the world better than its zero world if the 2 is replaced by a 3.) Although this principle seems plausible to us, we have not been able to prove that it preserves transitivity (nor that it does not). We are indebted to Lauwers for suggesting the modification used in the text so as to ensure the preservation of transitivity.

[15] We mention here a different sort of approach that agrees with many of SBI3's judgments. This approach was suggested to us by an anonymous editor of this JOURNAL, and was developed by Segerberg (his segmentation principle) in "A Neglected Family of Aggregation Problems in Ethics." Here is the principle: if (1) the locations of two worlds are the same, and (2) there is a partition of locations into regions such that w1 is at least as good as w2 in every region, but (3) not vice versa (that is, (2) with w1 and w2 permuted), then w1 is better than w2. For finitely additive theories, this is a plausible metaprinciple as far as it goes. This metaprinciple agrees with our metaprinciples in the examples discussed. But it is silent (inappropriately, we think) about whether <...−1,−1,−1,2,2,2,...> is better than its zero world. In contrast, SBI3 holds (appropriately) that it is better than its zero world. Note also that, unlike SBI3 (which evaluates larger and larger regions of worlds), the segmentation principle (which evaluates individual segments in isolation) would not be plausible for holistic theories of the good (such as strict egalitarianism).

Where the locations are the same in both worlds, SBI3 is the strongest metaprinciple that we shall defend. The following is an example where SBI3 rightfully remains silent:

Goodness at Locations (for example, times)
w1: ..., −1, −1, −1, 10, 1, 1, 1, ...
w2: ..., 0, 0, 0, 0, 0, 0, 0, ...
Example 6

SBI3 does *not* direct that w1 be judged better than w2. For although starting with the 10 location, all uniform expansions are better under w1 than under w2, starting with any ten adjacent −1 locations all uniform expansions are better under w2 than under w1. Thus, SBI3 is silent, and this, we claim, is appropriate.

Before defending SBI3, we shall introduce one more strengthening.

IV. AGGREGATING WHEN LOCATIONS HAVE AN ESSENTIAL NATURAL ORDER BUT ARE NOT THE SAME

So far, then, we have strengthened the basic idea to SBI3. SBI3 is applicable only where the locations of the two worlds are the same. It is silent if there is no meaningful basis for the transworld identification of locations, or if there is such a meaningful basis but the worlds do not in fact have the same locations. We shall now generalize SBI3 so as to cover a certain limited range of cases where the locations are not the same.

As indicated earlier, the mere existence of an essential natural order (a topological condition) does not guarantee the existence of an essential natural distance metric. If, however, there is an essential natural distance metric for locations, then we can say that two worlds are *isometric* just in case there is a 1-1 mapping of the locations of one world onto the locations of the other, such that the distance between any two locations in one world is the same as the distance between their mapped counterparts in the other world. (Such mappings will automatically preserve the topological structure.[18])

Where two worlds are isometric, we may appeal to *isometric counterpart functions*, which are simply distance-preserving 1-1 mappings of the locations of one world onto the locations of the another. Typically, there will be many distinct isometric counterpart functions between isometric worlds. If, for example, the locations have the metric structure of the real line, then $f(x) = x$, $g(x) = x+1$, $h(x) = x+2$, and so on, are all isometric counterpart functions.

---

[18] For simplicity, we are assuming that locations have a simple and uniform sort of structure. More specifically, we are assuming that (1) within a given world all locations have the same weight and are distributed uniformly by distance, if there is an essential natural distance metric; (2) all worlds have the same structure with respect to the topology, distance metric, and weights of locations. We are restricting our attention to such worlds so as to be able to focus on the core ideas. Things would be a lot more complex without these assumptions.

The idea, then, is that where two worlds are isometric but do not contain the same locations, some comparisons of the goodness of the worlds may be made by comparing isometric counterpart sets. For brevity, we shall formulate our metaprinciple in terms of *counterpart functions* where these are 1-1 mappings of the locations of one world onto the locations of the other (which, unlike isometric counterpart functions, need not preserve distances). Let us call a counterpart function *admissible* just in case (1) it is the identity function if the two worlds have the same locations, and (2) it is an isometric counterpart function if the two worlds are isometric but do not have the same locations.

Our generalized metaprinciple can finally be stated:

> GM (general metaprinciple): if (1) w1 and w2 have exactly the same locations, or are isometric, and (2) for all admissible counterpart functions, and for all bounded regions of locations in w1, there is a bounded uniform expansion and a positive number, k, such that, relative to all further bounded uniform expansions, w1 restricted to the expansion is k-better than w2 restricted to the counterpart expansion, then w1 is better than w2.[17]

To see the force of this generalization, consider once again example 4, repeated below:

Goodness at Locations (for example, times)
w1: ..., 5, 1, 5, 1, 5, 1, 5, 1, 5, ...
w2: ..., 3, 2, 3, 2, 3, 2, 3, 2, 3, ...
Example 4 (repeated)

If the locations are the same in each world, then GM, like SBI3, directs that finitely additive theories judge w1 as better than w2. For in this case, the only admissible counterpart functions are identity functions, and the force of GM is no different than that of SBI3. If, however, the locations are not the same, and there is no essential natural distance metric, then GM (like SBI3) is silent. In this case, the worlds are not isometric and there are no isometric counterparts. Except in certain special cases,[18] silence is appropriate, we claim, because without sameness of locations or isometry, there may be no way to rule out the possibility that w2 has the same locations as w1, plus a billion clones of each (or vice versa). And if this were the case, it would be a mistake to claim that finitely additive theories should judge w1 as better than w2.

---

[17] Throughout, we focus on metaprinciples governing when one world is *better* than a second. Each such metaprinciple has a natural companion metaprinciple governing when one world history is *equally good* as a second: one simply replaces all references to betterness with references to equal goodness. This will leave many worlds that are mutually incomparable (neither better than the other, nor equally good as the other). Alternatively, one could invoke a metaprinciple that judges two worlds equally good just in case neither is better than the other. This would have all worlds as comparable; but the relation of being equally good would not be transitive. We think the former approach more plausible, but shall not defend that view here.

If the locations are not the same, but there is an essential natural distance metric, then GM directs finitely additive theories to judge w1 as better than w2. For no matter what isometric counterpart function one considers (and there are infinitely many), relative to any uniform expansion of any bounded region, w1 restricted to that expansion is better than w2 restricted to the counterpart expansion.

We claim that GM represents a coherent and plausible way for finitely additive value theories to make judgments of goodness when aggregating over an infinite number of locations.

Before turning to a defense of GM, let us consider one final example that illustrates how GM has more power when the locations are the same than when they are merely isometric.

Goodness at Locations (for example, times)
w1: ..., −1, 0, 1, 2, 3, ...
w2: ..., −2, −1, 0, 1, 2, ...
Example 7

If the locations are the same, then GM (like BI) directs that finitely additive theories judge w1 as better than w2, since every single location is better under w1 than under w2. If the locations are merely isometric (not the same), then GM is silent. For although under the isometric counterpart function displayed every location in w1 is better than its counterpart in w2, under the isometric counterpart function that pairs the 0 location in w1 with the 1 location in w2 (and the location with *n* in w1 with the location with *n*+1 in w2), every location in w1 is worse than its counterpart in w2. Consequently, GM is silent if the locations are merely isometric. This silence, we claim, is appropriate.

---

[18] For simplicity, we shall not introduce two slight strengthenings of GM that are plausible. First, GM would remain plausible, we would argue, if it were revised to apply to worlds where all, except perhaps a finite number, of infinitely many locations are the same. For it is only where there are infinitely many nonshared locations that the problems of comparisons arise. Second, GM would remain plausible, if it were revised to apply also where the locations of one world are a subset of the locations in the other world. This is weaker in that it does not require that all locations be the same. It only requires that all the locations of one world (but not necessarily the other world) be in the other world. For finitely additive theories (but not for some types of nonadditive theories such as strict egalitarianism), this can be treated as if all the locations were the same in the two worlds by assigning "empty virtual locations" as counterparts for the locations are in one world but not in the other. For simplicity, we ignore these two strengthenings.

V. THE PLAUSIBILITY OF GM

In the previous section, we surveyed some examples and saw how GM applied to those cases. This discussion has served two purposes. One was simply to draw out the implications of GM. The other was to illustrate some cases in which GM gave answers that were plausible. We turn now to a general, and more abstract, defense of GM.

In assessing our defense of GM, two points should be kept in mind. The first is that there is nothing contradictory in combining GM with a finitely additive theory of value. By definition, finitely additive value theories are theories that evaluate worlds with a finite number of locations by adding up the total amount of local goodness. This is compatible with any sort of approach when there are an infinite number of locations.[19] It is typically implicitly assumed that the most natural and plausible way of evaluating goodness in such cases is on the basis of the total amount of local goodness. The examples of the previous section were intended in part, however, to provide evidence that, in order to be plausible, a finitely additive theory must assess infinite cases in a way that appeals to more than standard mathematical totals. GM is being defended as a plausible basis of assessment for such cases.

A second point to keep in mind is that GM only states sufficient conditions for one infinite aggregation to be better than a second. It does not state necessary conditions. We would be very surprised if it cannot be strengthened.

The justification of the plausibility of GM can instructively be given in two steps. The first is to defend BI, of which GM is a strengthening. The second is to defend the strengthening of BI to GM.

BI, recall, is:

> BI (basic idea): if w1 and w2 have exactly the same locations, and if, relative to any finite set of locations, w1 is better than w2, then w1 is better than w2.

---

[19] This same issue arises, of course, for probability functions. By definition, probability is finitely additive when the events are exclusive. But it need not be additive for an infinite number of exclusive events. A common view is that the probability that any particular event of an infinite number of exhaustive and mutually exclusive events will occur is zero, but the probability that at least one of them will occur is one. This is incompatible with infinite additivity.

KP-665

BI is a very weak metaprinciple. It has bite only if the locations are the same in the two worlds and every single location contains more goodness in one world than it does in the other.[20]

The plausibility of BI rests on the view that no new goodness considerations emerge for infinite sets when all finite subsets agree on what is better. Of course, this is not true as a matter of logic. There is no contradiction in aggregating things one way for all finite cases and differently for the infinite case. We are simply claiming that it is plausible to hold no new goodness considerations emerge in the above manner at the infinite level. Example 1 supports this view.

Now, probably the most powerful objection to BI is that it requires (at least sometimes) that merely shifting goodness from some locations to others can change the assessment of how good a world is. But surely, it might be insisted, mere shifting does not change anything. To see this, consider example 1 once again:

Goodness at Locations (for example, times)
w1: ..., 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
w2: ..., 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
Example 1 (repeated)

Assuming that the locations are the same, but without any other special assumptions, BI directs finitely additive theories to judge w2 better than w1.

The point to note here is that the distribution of goodness in w1 can be obtained from the distribution in w2 by shifting around some goodness (but not adding any). More specifically, w1's distribution of goodness by location can be obtained from w2's by picking an arbitrary location in w2 and shifting 1 unit of goodness to it from its right neighbor.[21] This yields 2 units of goodness in the selected location and 0 in its right neighbor. Then shift 2 (not just 1) units of goodness to the 0 location from its right neighbor. That yields 2 units in the formerly 0 location and −1 in its right neighbor. Then shift 3 (not just 2) units of goodness to the new −1 location, and so on. This leftward shifting process is repeated infinitely many times, and a similar, but rightward infinite shifting processes is done starting to the left of the original location. The net result is w1.

---

[20] For the purposes of expositional simplicity, we have ignored a metaprinciple that is weaker than BI. This weaker principle says that, if w1 and w2 have the same locations and every single location in w1 is better than every single location in w2, then w1 is better than w2. This metaprinciple would agree with BI in judging w1 as better than w2 in example 1 (where each location is 2-over-1), but it would not join BI in judging w1 better than w2 in a case where every location is either 2-over-1 or 4-over-3. For although each location is better in w1 (and thus BI judges w1 as better), some locations in w1 (namely, those containing 2 units of goodness) are worse than some locations in w2 (namely, those containing 3 units). Thus, the weaker principle is silent. In arguing that some sort of metaprinciple is plausible for distinguishing between some cases where two worlds have infinite value, the weaker metaprinciple is even more plausible (because weaker) than BI.

[21] We are not presupposing any natural order here. The order of the listing of locations may be arbitrary. 'Rightward' is to be understood as rightward relative to this ordering.

So w1 just is w2 with an infinite number of finite shifts of goodness. Surely, it might be claimed, BI is thus mistaken to judge w1 as better than w2. Since no new goodness has been added, surely w1 and w2 are equally good.[22]

BI is not mistaken. Note first that BI agrees that a finite number of shifts in goodness among locations does not change the goodness of a world (since a finite number of shifts will cancel out for any finite set that includes all the affected locations). It only holds that an infinite number of shifts can make a difference. When there are an unbounded infinite number of shifts, as illustrated above, one never has to settle one's accounts properly. One can keep borrowing unlimited amounts from the unbounded infinite pool of untapped creditors in order to give to others. Given this fact, w1 is indeed better than w2.

Another way of making this point is to note that the claim that infinite shifts of goodness do not change the goodness of the world is incompatible with a fundamental metaprinciple of finitely additive theories, namely, that, if the locations are the same in two worlds and every location (for example, every person) has more goodness in one world than in a second, then the first world is better. (For finitely additive theories, this metaprinciple is equivalent to BI.) Example 1 shows very clearly this incompatibility. For in this example every location is better in w1—even though w1 is "obtainable" from w2 by means of infinite shifts of goodness. Faced with the choice of rejecting this fundamental metaprinciple or recognizing that infinite shifts can make a difference in goodness, extensions of finitely additive theories will be more plausible if they do the latter.

---

[22] Note that a special case of shifting utility is permuting utility (that is, simply switching the utility at two or more locations). BI agrees that a finite number of binary permutations make no difference, but it allows that an infinite number of binary permutations can. For example, where the locations are the same and have a complete essential natural order (as in time), BI directs finitely additive theories to judge <1,1,1,0,1,0,1,...> as better than <1,0,1,0,1...>—even though the former is just an infinite permutation of the latter (namely, permute leftward all 1s in the latter that initially have a 0 to the left, then permute leftward again all 1s that still have a 0 to the left). Jorge Garcia and Mark Nelson object to this feature of the precursor of GM in their "The Problem of Endless Joy: Is Infinite Utility Too Much for Utilitarianism?" *Utilitas*, VI (1994): 183-92. Vallentyne replies in his "Infinite Utility and Temporal Neutrality," *Utilitas*, VI (1994): 193-99. In Luc Van Liederkerke, "Should Utilitarians Bother about an Infinite Future?" *Australasian Journal of Philosophy*, LXXIII (1995): 405-07, it is shown that the principle that infinite permutations (and shifts) make no difference in how things are ranked is incompatible with a weak Pareto principle (there called *monotonicity*). (The example just given is taken from that article and shows the incompatibility.) In Vallentyne, "Infinite Utility: Anonymity and Person-Centeredness," *Australasian Journal of Philosophy*, LXXIII (1995): 413-20, this result is used to support the view that infinite (but not finite) permutations can change how things are ranked (on the grounds that the Pareto principle is more compelling).

We conclude, then, that BI is plausible. We turn now to the question of whether our strengthening of BI to GM is plausible. Our most basic claim is that some sort of strengthening of BI is plausible. Again, all the examples of the previous sections other than example 1 illustrate this plausibility. Example 2 (in which all locations are 2-over-1 locations, except for one 1-over-2 location) is an especially compelling example. BI is silent here (since the singleton 1-over-2 set blocks the judgment by BI that w1 is better, and any 2-over-1 set blocks the judgment that w2 is better). But surely w1 is better.

Of course, GM may be mistaken in the manner in which it strengthens BI, but we think it is not. For GM is a strengthening of BI in four ways, and each is plausible. The first (starting with SBI1) is that GM requires only that every set of locations be such that, if expanded enough, relative to all further expansions (and all relevant counterpart functions), the goodness of the expanded set in one world is greater than that of its counterpart set in the other. This permits GM (unlike BI) to judge one world as better than another even if there are a finite number of locations at which it is worse. Example 2 illustrates the plausibility of this increased scope. The second strengthening (starting with SBI2) is that GM judges only bounded regional expansions as relevant. For locations with no essential natural order this changes nothing, since all expansions of a given set are bounded regional expansions (by definitional stipulation) in that case. For locations with an essential natural order, however, this is a strengthening, since it judges some logically possible expansions as irrelevant. And surely this is right. For the fact that there is an essential natural order means that the locations are ontologically fixed in a certain order, and therefore that certain logically possible ways of expanding a given set of locations are incompatible with this natural order. Expansions that violate this natural order can thus be ignored. The third strengthening of GM over BI (starting with SBI3) is that only uniform regional expansions are deemed relevant. Roughly, these are regional expansions that expand in all directions by the same distance (if there is a distance metric) or by the same number of locations (if there is no distance metric but locations are discrete). The restriction to uniform regional expansions is plausible because it rules out "nonrepresentative" expansions, which expand along some dimensions but not others (and thus fail to represent the structure of values of the world as a whole). The fourth strengthening is that GM is applicable where the locations are isometric, but not identical. Isometric worlds have exactly the same distance relations: the distance between any two locations in one world is exactly the same as the distance between their counterparts in the other. Consequently, it is plausible to make judgments whenever all isometric counterpart functions yield the same judgment. We thus conclude that both BI and the stronger GM are plausible.

VI. CONCLUSION

Finitely additive theories of value rank worlds with a finite number of locations (peoples, times, and so on) on the basis of the total goodness they contain. It is commonly supposed that the only way that such theories can apply to worlds with unbounded infinite numbers of locations is to rank them on the basis of the total goodness they contain. But this has the crazy result that where time (for example) is unbounded, a world with 2 units of goodness at each time is not better than a world with the same locations but only 1 unit at each location. Fortunately, this way of dealing with the infinite case is not necessary for finitely additive theories. And given the implausible rankings it generates, it is not plausible.

We have argued in favor of a different approach. At the most basic level we have defended BI, which says that, if two worlds have the same locations, and, relative to every finite set of locations (and a fortiori, every location), one world is better, then that world is better tout court. Somewhat more tentatively, we have defended the strengthening of BI to GM. GM is a strengthening in that, as long as the two worlds are location isometric, it applies even when the locations are not the same in the two worlds. Furthermore, it does not require that one world be better than the other, relative to all finite sets of locations, but only to all bounded uniform expansions of some bounded uniform expansion of any bounded region of locations. Although we can imagine that, due to a faulty understanding of the issues, we may be mistaken about the plausibility of these strengthenings, we are very confident that some strengthenings of BI are plausible. And it should be remembered that we are only claiming that GM is plausible as far as it goes. We certainly do not think that GM is the strongest plausible metaprinciple for dealing with unbounded locations.

Finally, in closing we note that, although we have only defended GM as a plausible metaprinciple for finitely additive value theories, it is also probably plausible for other kinds of aggregative value theory. For it makes no assumption about how bounded sets of locations are evaluated. It simply says that when a judgment of betterness holds for all bounded sets of a certain sort, then that judgment holds for the entire unbounded set as well. We think, for example, that GM is plausible for maximin theories of value (which evaluate the goodness of a world on the basis of the minimum goodness at any location), for average theories (which evaluate the goodness of a world on the basis of the average goodness at locations), and for strict egalitarian theories (which evaluate the goodness of a world on the basis of how equally goodness is distributed). But seeing how far GM is applicable to these and other theories of value is something that must await another occasion.

PETER VALLENTYNE

Virginia Commonwealth University

SHELLY KAGAN

Yale University

REVIEW ESSAYS

Socratic Studies. GREGORY VLASTOS. MYLES BURNYEAT, editor. New York: Cambridge University Press, 1994. xi + 152 p. Paper $16.95.*

Gregory Vlastos died in October 1991 at the age of eighty-four. He had completed for publication the important volume, *Socrates: Ironist and Moral Philosopher*,[1] and he had planned a companion volume to appear in conjunction with it (*ibid.*, pp. 18-19). The companion work was to contain revised versions of three previously published essays, plus some new material: "Sadly," writes Myles Burnyeat, the volume's editor, "not as much new material as he had planned to write" (ix). The previously published essays are: "The Socratic Elenchus" and "Socrates' Disavowal of Knowledge," both extensively revised, and "The Historical Socrates and Athenian Democracy," which appears with minor revisions. The new material consists of an article on Socratic definition entitled "Is the 'Socratic Fallacy Socratic'?", an article on the accounts of courage in the *Protagoras* and *Laches*, completed from drafts and notes by Burnyeat and others, and a speech given to graduates in classics at Berkeley in 1987, entitled "Socrates and Vietnam." Only "Socrates and Vietnam" was inserted without the explicit authorization of Vlastos; Burnyeat chose it as a fitting counterpart to the epilogue, "Felix Socrates," in the 1991 book.

Vlastos's rationale for separating the two volumes was never fully clear. He apparently wanted the 1991 book to pursue a connected line of argument about the relationship between Socrates and Plato, without being sidetracked by detailed considerations of method, epistemology, and history that might prove of interest only to more specialized scholars. Nonetheless, his accounts of the elenchus, of Socratic ignorance, of the relationship between mathematics and the search for definition, and of Socrates' relationship to the Athenian democracy all figured (briefly summarized) in the argument of *Socrates*, providing crucial props in Vlastos's argument dividing Socrates from Plato. This volume, then, is an indispensable companion to its predecessor.

I

One of the major achievements of Vlastos's career is his analysis of the Socratic elenchus, the procedure by which Socrates subjects the beliefs of his interlocutors to scrutiny. Socrates famously questions

---

* I am grateful to Richard Kraut, Richard Posner, and Cass Sunstein for comments on an earlier draft of this review.

[1] Ithaca: Cornell, 1991.

0022-362X/97/9401/27-45

rather than asserts; he himself would seem to make no contribution to the unfolding inquiry. And yet he appears to claim that moral progress has been made. In the *Gorgias*, he even claims to have "proven" moral truths by elenctic methods, even that the conclusions "have been clamped down and bound by arguments of iron and adamant" (508E-509A). It is not surprising that such claims should continue to fascinate moral philosophers, especially in the contemporary atmosphere of skepticism about transcendent grounds for moral understanding. It is natural to ask whether, if inquiry can only be the internal probing of human belief for consistency and explanatory soundness, any progress toward truth can really be made. One reason why Socrates' method continues to fascinate is that it appears to promise an affirmative answer to this question. For this reason, Vlastos's thorough exploration of the method has been of interest to many nonspecialists as well as to specialists.

Vlastos defines the elenchus as follows:

> Socratic elenchus is a search for moral truth by question-and-answer adversary argument in which a thesis is debated only if asserted as the answerer's own belief and is regarded as refuted only if its negation is deduced from his own beliefs (4).

Vlastos insists that the elenchus must be regarded as a search for truth, since only this aspiration would serve, in Socrates' mind, to distinguish it from "eristic," contentious argument aimed at point scoring or defeat of an opponent. He rightly emphasizes that Socrates consistently depicts his mission as involving a positive search of some type, not simply as disputation for its own sake.

Another valuable contribution is Vlastos's stress on the "sincerity condition"—the repeated insistence of Socrates that the interlocutor say what he really believes, not just any old proposition that might help him win the argument. It is only on this basis, Vlastos rightly argues, that Socrates can regard the elenchus, as he plainly does, as a way of scrutinizing a person's entire life, calling its fundamental commitments into question. The elenchus has, in Vlastos's words, "a double objective: to discover how every human being ought to live *and* to test that single human being who is doing the answering—to find out if *he* is living as one ought to live"(10). These two objectives can only be pursued together if the interlocutor produces sincere answers and thus exposes his inner world to scrutiny.

Vlastos contributes little that is new to the logical analysis of the elenchus. Typically, as he states, the interlocutor asserts a thesis, p. Socrates then secures agreement to further premises, say, q and r.

Socrates then argues, and the interlocutor agrees, that *q* and *r* entail not-*p*. In many cases, Socrates stops there, allowing the interlocutor and the reader to draw their own conclusions. They are at liberty to reject any of the premises, or to find equivocation or fallacious inference in the argument. It was Vlastos's earlier view[2] that Socrates' enterprise went no further than this: his aim was simply to reveal to his interlocutors their confusions and muddles, stimulating further searching and destroying the "conceit of wisdom," to use the phrase from the *Apology* (*Ap.* 21B-23B). This was also George Grote's interpretation, as Vlastos shows.

This picture of Socrates, however, later seemed incomplete to Vlastos. If that was all Socrates expected to get out of the elenchus, where did he find positive support for the views about virtue and goodness on which he based his life, and recommended to others? For Grote, the negative cross-examination and the "affirmative dogmatism" were simply separate aspects of Socrates' career; they did not even mutually support one another. Unwilling to accept this result, in which the missionary of the examined life turns out to be a dogmatist himself, Vlastos began to notice that there was textual evidence against it. Crucially, in the *Gorgias* he found Socrates repeatedly asserting that his elenctic argument has "proven" his thesis true and his interlocutor's false. Socrates maintains that when he argues against an interlocutor, he "knows how" to make this interlocutor bear witness against his own proposition (464A); conversely, he insists that the truth "is never refuted" (473B).

How can Socrates believe this, Vlastos asks? How do we get from a display of inconsistency to the proof of a truth? And, once the inconsistency is laid on the table, why should we think that the choice of which proposition to throw out will be done the right way rather than the wrong way? Vlastos's bold answer to this problem, which, in effect, is *the* problem of elenctic argument both now and formerly, is to ascribe to Socrates the confident belief "that if that wrong choice were made *he would have the resources to recoup that loss in a further elenchus*" (22). This, in turn, Socrates can hold because he holds the more general belief *A*:

> (*A*) Whoever has a false moral belief will always have at the same time true beliefs entailing the negation of that false belief.

The idea is that, if this is so, and if elenchus can go on long enough and range widely enough, then the only consistent set will be the one including true beliefs.

---

[2] See 17, citing his introduction to the *Protagoras*, and his essay "Paradox of Socrates," in Vlastos, ed., *The Philosophy of Socrates: A Collection of Critical Essays* (Garden City: Doubleday, 1971), pp. 1–2.

There are two large problems with this claim. First, Socrates gives us no reason to suppose that it is true. Given the wide range of interlocutors examined by Socrates, and given his own views about the corruption of values and morals in Athenian society, how could he feel confident that the true beliefs about moral value are in there along with the false? Second, thesis *A* gives us reason to think only that the truth will prevail over the very long haul, not that it will prevail or has prevailed in any particular argument. It thus provides no support for Socrates' contention that he has established any particular thesis by elenctic argument.

Vlastos's answer to the first problem is that it is indeed a problem for Socratic procedure and that Plato sees it as such. He suggests that Plato's theory of recollection is his own answer to the problem: we can safely assume that all souls contain the truth only if we believe that they have been connected to the truth in some reliable manner before they enter this world of error. Thus, the philosophical problem itself becomes a confirmation of the textual interpretation. Although this is an attractive feature of Vlastos's proposal, it would be useful if we could know more about what Socrates himself thought about the origins of truth in democratic man.[3] It is peculiar that Vlastos says so little about this, given his evident admiration for Socrates and his suspiciousness of Plato's other-worldly and elitist views about moral knowledge.

More troubling in interpretive terms is the claim that Socrates, relying on *A*, has proven certain particular propositions. For we want to know why Socrates has any reason to be confident that in this particular case the truth has triumphed over error. In a set of incisive comments on Vlastos's original article, Richard Kraut[4] makes a helpful suggestion: Socrates can have confidence in many of his premises because certain beliefs are so eminently reasonable that they do not need justification *as yet*, or unless they are challenged. Such a reading of Socrates agrees with Vlastos in rejecting any appeal to consensus or self-evidence to anchor the elenchus; but it at least gives us a place to start in assessing the likelihood that a particular example of elenctic argument has established its conclusion, since it gives us at least some reason to trust the premises of elenctic argument.

---

[3] In "Socratic Puzzles," *Phronesis*, XL (1994): 143-55, Robert Nozick suggests that the elenchus is reliable toward the truth only when a person *knows* truths about a subject matter sufficient to introduce inconsistency with any false belief he holds, and proposes revising Vlastos's view along these lines.

[4] "Comments on Gregory Vlastos, 'The Socratic Elenchus'," *Oxford Studies in Ancient Philosophy*, I (1983): 59-70.

As Kraut and Vlastos agree, however, this is a starting point only, not a full solution to the problem. To explain fully Socrates' confidence in his "proofs" of highly controversial moral propositions, Vlastos ascribes to him a further strong assumption, *B*:

> (*B*): The set of elenctically tested moral beliefs held by Socrates at any given time is consistent.

B and *A* together entail that Socrates' belief set consists of entirely true beliefs, and this could certainly explain his confidence that an argument proving the truth of one of these beliefs and the falsity of its contradictory is a successful piece of proof.

But at a cost: for now we have a Socrates who is confident that a procedure has established something only because that something has been established already by his own process of self-inquiry. The arguments in the dialogues in effect show other people from their own belief set what Socrates knows already from his, and elenctic inquiry is not really inquiry, in the sense that the truth is antecedently established. We have a Socrates who is more complacent and less puzzled than the dialogues indicate. (Kraut rightly points to evidence of perplexity and tension within Socrates' own belief system.[8]) I shall return to this issue, since it appears to be a source of some inconsistency in Vlastos's own portrait of Socrates.

Vlastos's interpretive thesis is elaborated in an afterword in which he stresses the uniqueness of the *Gorgias* among the Socratic dialogues: only here does Socrates claim to have proven his conclusions, and only here do we have firm grounds for ascribing to Socrates theses *A* and *B*. Vlastos now suggests that Socrates was committed to *A* and *B* throughout his procedure, but unawares. Not being an epistemologist, he went about his business of scrutiny without ever consciously formulating its presuppositions. Thus, he was not aware of the controversial nature of his own strong assumptions. It was his pupil Plato who, in the period when he wrote the *Gorgias*, had become engrossed in epistemology and was therefore increasingly restive about the elenchus. Although in earlier works he had recorded his teacher's activity without bringing its epistemological presuppositions to the surface, in the *Gorgias* they become clear to him and he records them. Having recorded them, he sees how "frightfully strong" they are, and begins to feel that it would be "hopeless" to justify them by the sort of evidence Socrates could have offered. The result is that Plato loses faith in the elenchus. Immersing himself in mathematical study, he arrives at an altogether different picture of justification, which supercedes elenchus in the middle dialogues, a picture involving recollection and quasi-geometrical deduction.

---

[8] Kraut, p. 69, discussing passages from the *Protagoras*, *Hippias Minor*, and *Gorgias*.

Once again, Vlastos tries to turn a philosophical problem into an interpretive advantage, by producing a thesis that makes the philosophical impasse a crucial turning point in Plato's philosophical development. The interpretive conjecture is plausible, though far from certain. But it leaves some important issues unresolved, both interpretive and philosophical. In terms of our understanding of the texts, we want to know just how this reading of the *Gorgias* helps us in understanding the earlier dialogues, where Socrates neither states anything like assumption *A* nor claims to have proven the truth of propositions. Vlastos would have us understand that he does hold *A*, if not self-consciously, and does take himself to be proving the truth of his conclusions therewith. If so, there is still a puzzle as to why he never claims to have proven truths, and indeed often fails to indicate what solution to the elenctic impasse he himself favors. Philosophically, we are left with a large worry. Since Vlastos is plainly far more sympathetic to Socrates than he is to the mature Plato, and plainly finds the elenchus very attractive as a philosophical method, how would *he* propose to resolve the problems posed by his own analysis? If not in Plato's way, by turning away from elenchus altogether, then how?

II

But we cannot go further on these issues without introducing another piece to the puzzle. In "Socrates' Disavowal of Knowledge," Vlastos tackles another interpretive problem of long standing. How are we to understand the fact that Socrates professes ignorance, not just of definitions,[6] not just of particular moral propositions, but ignorance quite generally, while at the same time claiming knowledge of particular moral claims, such as the fact that injustice is base (*Ap.* 29B6-7) and the fact that a good man cannot be harmed[7] (*Ap.* 41D). In the *Gorgias*, he actually combines a strong assertion that he has proven his conclusion "by arguments of iron and adamant" with the claim that "I do not know how these things are" (508E-9A). Vlastos argues, plausibly, that Socrates is implicitly contrasting a type of human, still fallible knowledge available through the elenchus with a stronger totally infallible type of knowledge that is, in his view, unavailable to finite human beings. The distance between the two types of knowledge is understood in religious more than in analytic terms: his disavowal is a way of conforming to a traditional piety that urges mortals to "think mortal thoughts."

---

[6] Nozick resolves the puzzle by ascribing to Socrates ignorance only of the answer to the 'What is *x*?' questions on which he focuses; Socrates can, he argues, be ignorant of such definitional questions while knowing a number of more concrete moral truths. Unfortunately, the texts do not support this claim: Socrates claims he lacks wisdom "in anything, great or small" (*Ap.* 21B2-5), denies knowing "anything noble or good" (21D2-6), and so forth. The denial is perfectly general.

[7] Here, Socrates does not say that he "knows" the truth of the thesis, but he strongly urges his audience to "apprehend this one thing as true," which at the very least entails confident true belief. As we shall see, Vlastos's distinction between elenctic knowledge and certain knowledge is not the same as the distinction between true belief and knowledge, since elenctic knowledge admits the possible falsity of the proposition.

On this view, does Socrates claim to have true belief, while simply denying that his beliefs meet some further condition that would suffice to turn true belief into knowledge? Or is he worried that his beliefs might after all not be true? Vlastos appears to choose the second option, when he writes that a proposition *Q* that has turned out true in a thousand elenchi "might still turn out false in the thousand-and-first." Furthermore, he adds, Socrates' victory in all arguments may simply show "that he is the better debater. It could not show that there is no inconsistency within his own set of beliefs—only that, if there is, no opponent has managed to spot it" (57). Finally, he observes, elenctic "knowledge" is "full of gaps, unanswered questions; it is surrounded and invaded by unresolved perplexity" (58). All of these reasons seem to tell in favor of not claiming for sure that his beliefs are true,[8] though at the same time he may confidently assert them as beliefs that have passed muster in the best tests he and others have devised. In that sense, in a more colloquial sense of 'know', he may be said to "know" them. To use the less misleading language of *Apology* 41D, he can strongly urge his audience to "apprehend [them] as true," without ruling out the possibility that someone someday may show them to be false.

Robert Nozick, commenting on this problem, has offered a useful example of how a philosophical thesis may pass all our toughest tests in such a way that we may confidently assert it, yet at the same time lack full certainty that it is true. Think of the history of the attempt to define "knowledge," he says. Each definition, over time, has proven vulnerable to some counterexample that was not foreseen, and sometimes took centuries to produce. Nozick thinks his own definition true. He confidently asserts it. In some informal moment, he might claim to know it.[9] He certainly invites his audience to apprehend it as true. It passes all our tests, and it can also help us to see how all previous accounts failed. On the other hand, it is reasonable, given the history, to believe that there may well be some unseen flaw in the definition that may send it the way of its predecessors. It would still have been closer to the truth than they, but it would turn not to be really true. So, too, he suggests, Socrates feels that his own theses may fall[10] to unforeseen objections. And yet he asserts them as the best we currently have.

---

[8] Vlastos presumably means that Socrates thinks the answers to one of the unanswered questions might conflict with beliefs already in place and force us to reexamine them. (Compare John Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971), pp. 20, 48-51, on the way in which an inquiry into principles may at times cause us to revise our considered judgments.) But note that, if Socrates really demands that all gaps be plugged before any knowledge claim can be entered, he is making comprehensive understanding a prerequisite of knowledge, an unduly strong demand.

This offers an attractive and plausible account of the Socratic disavowal. If it does require us to charge Socrates with using a central term in two misleadingly different ways,[11] we may still say that the use of ambiguity in a deliberate paradox is not objectionable in the usual way, since the point of a paradox is to stimulate reflection. Socrates' paradoxical double claim has amply justified his expectation that this combination of propositions, instead of tricking people into false veneration for Socrates, would get them thinking about what moral knowledge really is.

The real difficulty with this thesis of Vlastos's, it seems to me, is not its relation to the texts or its own intrinsic plausibility. It is its relation to Vlastos's account of the elenchus. For how can Socrates hold the views imputed to him here, concerning the possibility that his own views will come unstuck, and at the same time hold propositions *A* and *B*? Vlastos does not comment on this issue, but I think there is a way in which he could solve the problem, while at the same time solving a major problem in his interpretation of elenchus, namely, its excessively strong reliance on texts in the *Gorgias*.

We might ascribe to Socrates the following view: all people hold beliefs that, in the fullness of time and with sufficient patience, will be shown to entail the falsity of any false beliefs they may assert in elenctic argument (that is, *A*, so interpreted). Furthermore, (*B*) the set of beliefs tested and held by Socrates is as good a candidate as we currently have for truth, and so it is a prima facie strike against a belief that it conflicts with a member of this set. (*C*) We do not really *know* infallibly that Socrates' beliefs are true, however, or that some future elenchus will not bring that whole edifice crashing down. Socrates' beliefs conform to a human standard of elenctic knowledge, and so it makes sense for us to proceed on the hypothesis that they are true; but that leaves open the possibility that they might actually be false.

---

[9] Nozick does not say this, because he thinks that Socrates' knowledge claims are not about the same things as his professions of ignorance. Since I have rejected that part of his thesis, I am adjusting his example to make the parallel with Socrates, as I understand Socrates' position, more precise.

[10] Nozick puts the point more strongly—Socrates believes it is likely that his own theses "would" be shown to be false. This seems unnecessary, and hard to square with the texts.

[11] A central worry of Nozick, p. 144.

This modest view makes good sense of what we find in the Socratic dialogues with the exception of the *Gorgias*. It tells us how Socrates can combine a positive moral search with humility about the actual results of human efforts to know. What we can then say (following Vlastos's own lead) is that in the *Gorgias*, Plato, dissatisfied with such a modest and human undertaking as that, depicts a Socrates who really delivers proofs and justifications of particular theses, not just the more modest human claims of the earlier Socrates. He shows us that this Socrates would have to believe both *A* and *B*, and that these conditions are extremely strong. This suffices to engender in Plato (perhaps in Plato's reader) the idea that the elenchus cannot plausibly do the job moral inquiry needs to do, and that some other method must be found. In other words, we attribute to Plato not a conscious formulation of what Socrates previously has assumed without formulating, but, instead, a subtle shift and a sharpening in the entire project, as the idea of proof, so central to the mathematics of which he was so increasingly enamored, comes to dominate his picture of ethical inquiry. (We do not need to follow Vlastos in holding that Plato now abandons the elenchus completely, a conclusion that appeals to Vlastos only because for other reasons he holds that *Republic* I was written before the rest of the work, as a separate Socratic dialogue. If we do not believe this, we should conclude that Plato continues to use the elenchus, but supplements it with other methodologies inspired by mathematics.)

What now of the philosophical project? What should Socrates' contemporary readers think about the possibilities of elenchus? They should say, I think, that such a method may continue to yield the human and fallible kind of result Socrates claims for it, but that it will do so most reliably if we do some further work that Socrates did not do.

First, if we are going to defend a thesis like *A*, we are going to need to give an account of human beings, and of the project of ethical inquiry, that will explain why we should suppose that each and every one of them, including the most corrupt, has at least some true beliefs. Here we can proceed in two different ways. We may decide to weaken *A* itself, by insisting that we are interested not in all actual people, but in people who meet certain minimal conditions of reasonableness; then we shall have to show how, despite disagreements among these people, it is nonetheless plausible to suppose that they all have some true beliefs. (Such, for example, seems to be the way in which John Rawls has adapted his Socratic project, in recent work, to the task of political justification.) Or we can decide to keep the elenchus universal, and still argue that there are reasons to hold that each and every human being has some true moral beliefs. If we take the latter course—or even the former—we shall probably point to the fact that the problems dealt with by moral theory derive from certain commonalities of human experience, such as the badness of pain and hunger and death, and that these supply some universal starting points that can be expected to be shared even by people who may hold all sorts of other beliefs that are false and/or inconsistent. We will want at the same time to advance an account of moral development, in order to show how it happens that human beings do form reliable beliefs about these basic matters. On all of this, Socrates seems to have little to offer us; we shall do better to turn to philosophers who have understood the importance of confronting these background questions: for example, Jean-Jacques Rousseau and Rawls.

Then, second, if we want to use the elenchus to establish reliable conclusions in particular finite chains of argument, we had better have some way of weighting the reliability of premises and judgments. Kraut seems right that Socrates implicitly tackles this problem, but since he offers no explicit criteria for reliability, he does not help us get on with it. We can get help here from Aristotle (a philosopher of whom Vlastos is unjustly dismissive, because he holds, wrongly, that Aristotle's method makes it impossible to establish unconventional ethical conclusions); Aristotle's conception of the person of practical wisdom and the nature of his experienced judgments gives us at least a start on the issue. A contemporary project, explicitly traced to both Socrates and Aristotle, is Rawls's account of ethical justification, with its use of a notion of "considered judgment" and its criteria for such judgments, which supply "provisional fixed points" in a moral inquiry.[12]

Finally, if we are to have any hope of using elenchus to get to truth in a corrupt society, we need an account of the likely sources of distortion in our socially learned value judgments and in the emotions that incorporate them. Aristotle had something to say about the likely biases toward money and honor in his own society, but it was the Hellenistic philosophers who developed this part of the Socratic project in a philosophically useful way, attempting to show how even a corrupt Roman interlocutor may be led by argument to see the inconsistency of certain Roman norms with more deeply held notions of human agency and dignity. Whether those projects are successful, and whether, insofar as they are successful, they are compatible with Socratic methods is, of course, another story. What is clear is that Socratic argument needs an account of vice—obviously one that will itself be tested and retested during the elenctic process—if it is to get off the ground in the sort of society that tried and condemned Socrates.

---

[12] See *A Theory of Justice*, pp. 20ff. and 47ff.

In short, Vlastos has shown that the elenchus can still command our interest. He also shows that it has many gaps and shortcomings, gaps that ultimately led Plato to reject it, or at the very least to judge that it needs to be supplemented by other methodologies. This history invites those of us who still think the elenchus our best hope for moral justification in an imperfect world to scrutinize it and test it further, doing the work which Socrates did not yet do and which Plato thought not worth doing.

III

In "Is the 'Socratic Fallacy' Socratic?" Vlastos returns to the distinction between two types of knowledge, using it to solve a famous problem. Peter Geach[13] ascribed to Socrates a problematic thesis, which he branded a fallacy. Vlastos rephrases the thesis as follows:

> (*G*) if you do not know what the *F* is, you will not know if you are predicating "*F*" correctly about *anything* whatever—you will not know if *anything* is *F* (69).

The unqualified generality of the claim is the problem: if it were only the claim that in the absence of knowledge of the definition you will lack knowledge of some case or cases, the claim would be reasonable enough.

Vlastos begins by noting that the evidence for attributing *G* to Socrates comes from the *Lysis* and *Hippias Major*, dialogues written, in Vlastos's plausible view, relatively late in the Socratic group. He calls them "post-elenctic dialogues," meaning that by this time Plato is losing faith in the elenchus and searching for other types of argument that will yield the moral certainty he requires. Vlastos now asks what the Socrates of the early dialogues would have made of *G*. If we introduce the distinction between certain knowledge and elenctic knowledge, he argues, we get a clear result: "*G* turns out to be a paper tiger" (76). If we are talking about fallible elenctic knowledge, Socrates thinks we certainly can have that kind of knowledge of particular instances independently of knowledge of the definition. He routinely treats our grasp of many particular examples of *F* as secure, and as providing a basis on which to build our search for the definition. In texts such as *Euthyphro* 6E, he does state that knowledge of the definition is a sufficient condition for secure knowledge of all instances, but he never makes it a necessary condition for knowledge of any. On the other hand, certain infallible knowledge of instances would, like all transhuman knowledge, have been of little interest to Socrates, who defines his moral quest in terms of the sort of knowledge that is available to a human being. It is only because Plato, dissatisfied with what the elenchus can offer, makes a stiffer set of demands that *G* comes to bother him. "[U]nder the spell of his own deepening mathematical interests," Plato comes to think of definition as crucial for success in the search for infallible knowledge; this prompts still further dissatisfaction with Socratic methods, and, ultimately, the *Republic*'s search for a completely unhypothetical knowledge of definition.

---

[13] "Plato's *Euthyphro*: An Analysis and Commentary," *Monist*, L (1966): 369-82.

Once again, this is on the whole an attractive story about the relationship between Socrates and Plato. Once again, however, it fails to grapple squarely with the question how Socrates can really have secure knowledge, even elenctic knowledge, of any concrete proposition, given the finite nature of any particular elenctic exercise. Once again, the analysis leaves us longing for a fuller account of the judgments in which we should have confidence during the elenctic process. Why should our grasp of examples be taken by Socrates as more secure than our grasp of definitions, as it surely is? What is it about our relationship to particular instances of *F* which is more favorable epistemologically than our relationship to the general account? And is Socrates always correct to prefer the cases to the general account? In *Republic* I, for example, a Kantian would say that Cephalus' definition of justice is on a sounder footing than Socrates' counterexample: it is in fact always just to tell the truth and pay back what you owe, even when the madman asks for the return of his weapon. Vlastos's account, elegant as it is, ultimately disappoints by its failure to probe more deeply into these philosophical issues.

IV

In "The Historical Socrates," Vlastos argues for two theses. (1) In his own time, Socrates was widely perceived as an antidemocrat.

But (2) this public perception was a misperception: though critical of the conduct of public life at Athens, he preferred democracy to any other constitution. Since (1) is uncontroversial, most of the argument concentrates on (2). Vlastos finds a few shreds of external evidence: Socrates' long association with the strong democrat Chaerophon, a leader of anti-oligarchic resistance in 404 B.C., and the continued admiration of Lysias, a strong democrat who defended Socrates in public after his death. Vlastos argues plausibly that Lysias would hardly have defended a man whom he took to be in league with the oligarchic cause he detested.

Most of the argument, however, rests on texts: and here we are confronted by a contradiction. Xenophon, himself pro-oligarchic, portrays Socrates as a crypto-oligarch, disdainful of the people and of filling offices by lot, opposed to equal citizenship for craftsmen, whose sedentary occupations are said to make their bodies and characters "womanish." Xenophon's Socrates argues that only expertise confers legitimacy on a ruler; this expertise is not primarily moral, since knowledge of ends is taken to be a relatively straightforward business. The hard work concerns itself with the instrumentalities of a good life: finance, diplomacy, civic harmony. Plato's Socrates, by contrast, portrays the expertise relevant to politics as moral expertise, knowledge of the ends of the good life. Moral self-examination, as he both taught and showed, is open to every human being, not just to the elite whom Vlastos felicitously calls "the beautiful people" (rendering *kaloi kagathoi*), and is in fact a necessary part of a worthwhile life for every human being. "The call 'perfect your soul' goes out to everyone, delivered personally by Plato's Socrates to 'anyone he runs into' in the market or on the street" (103).

But what exactly does this have to do with political democracy? One may certainly hold Socrates' views about the examined life while denying, for one or another reason, that democracy is the best regime, or even while maintaining that the question of the regime is not very important. The Greek and Roman Stoics, who followed Socrates in his democratic call to self-examination for all, extending it to male and female, slave and free, were in politics many things: republicans, aristocrats, defenders of empire, or even rulers of empire. None was a direct democrat, Athenian style. Certainly not all conceived their egalitarianism about women's and slaves' self-examination as having radical consequences for political citizenship or even the institution of slavery.[14] So Vlastos needs more evidence for Socrates' political ideas. He finds it in the speech of the Nomoi of Athens in the *Crito*.

The Nomoi are usually called "the Laws"; but we should remember that the Greek term *nomos* has a far wider extension, taking in customs, habits, and ways of life, as well as formal laws and constitutional structures. When the Nomoi come to Socrates, they offer him "strong proof that both we and the city have been pleasing to you" (52B). The "proof" is that Socrates has stayed in the city and has left it far less than most people. This seems to the Nomoi evidence of an intense preference for them and for civic life in accordance with them. This preference, they go on to argue, exacts obligations now.

What, says Vlastos, keeps Socrates in Athens "like an infatuated lover"? Only *nomoi* are mentioned, along with the city itself. What is it that makes Socrates prefer Athens to "every other city, Greek or barbarian," including some that are, by his own account, "well governed" in comparison to Athens? Vlastos reasons that it cannot be the civil or criminal law, since Athens differed little from its neighbors in these respects. He then concludes without entertaining further options that it must be political democracy that draws Socrates' love—a proposition, note, for which the text gives no evidence at all.

Can we do more either to support Vlastos's hypothesis or to cast doubt upon it? Surely, the *Crito* itself casts at least some doubt: for at 45A2-D5, Socrates argues that in moral decisions, as in medical decisions, it is the expert we should trust. Does he believe that the requisite expertise can be cultivated in a democracy? Apparently not: for at 49C10-D5 he says that most men will always reject the principle that injustice must never be done.[15] Perhaps, in keeping with Socrates' views elsewhere, we could hold that the requisite expertise can be cultivated in a democracy if citizens will subject their beliefs to the elenchus and continue to do so throughout their lives. But, of course, Socrates' present predicament is good evidence that up until now the Athenian democracy has not been friendly to that idea.

---

[14] The Greek Stoic ideal city did admit women as citizens on fully equal terms: see M. Schofield, *The Stoic Idea of the City* (New York: Cambridge, 1991); but Musonius Rufus, while defending women's philosophizing as necessary for good life, imagines that they will use their Socratic practice to enrich the home. Seneca (*Moral Epistle* 47) argues that the good Stoic should treat his slaves humanely and respect their humanity; but he clearly thinks that this requires only reform of the institution (for example, by consulting more with slaves, and by not using them as sexual tools), not its abolition. On slavery and Roman Stoic attitudes, see Miriam Griffin, *Seneca: A Philosopher in Politics* (New York: Oxford, 1976), pp. 256-85.

[15] See the good treatment of this material in Kraut, *Socrates and the State* (Princeton: University Press, 1984), p. 244.

We might wonder, then, whether there is some candidate for the object of Socrates' preference that Vlastos has not considered, something that Socrates loves and prefers about Athens even though it fails to meet his demand for moral knowledge or even serious self-examination. If we consider the breadth of application of the term '*nomoi*', we have a powerful alternative: the ways of life in Athens that give latitude for and encouragement to the intellectual life, the *nomoi* that Thucydides' Pericles praised when he said: "We pursue our love of beauty without extravagance, and our love of wisdom (*philosophoumen*) without softness." Socrates might have thought Sparta's the superior political order on the whole, while preferring Athens for all the things it had in abundance: art, philosophy, the freedom to speak, a vibrant intellectual culture. Sparta had none of these things, nor, to the same degree, did any other city. If we have to plump for just one of these items as the decisive object of his preference, we should surely focus on the freedom of expression—for this was a conspicuous feature of Athens in contrast to her neighbors,[16] and it is a feature highly relevant to Socrates' philosophical mission. Socrates might well believe that Spartan repression of this freedom undercut the opportunity of individuals to lead "the examined life," and therefore militated against human well-being,[17] without thinking this feature intrinsically bound up with a democratic constitution, especially one of a direct majoritarian kind. (Indeed, as his conviction shows, majoritarian democracy gravely threatened this freedom.)

To show Socrates that a general preference for Athenian freedom entails a preference for democracy as the best constitutional order, the Nomoi would have to construct an argument of the type attempted by Thucydides' Pericles, according to which only democratic self-government permits the appealing *nomoi* of free expression to flourish. The Nomoi attempt no such Periclean argument. They would have had an uphill task, since Pericles is among Socrates' least favorite politicians,[18] and the funeral oration comes in for an (albeit mysterious) skewering in the *Menexenus*. Perhaps, however, they might have persuaded him at least that the well-ordered regimes of Sparta and Crete are bound to inhibit the critical self-examination that is requisite for virtue; thus, we should put up with democracy despite its faults, since the faults of other systems are ultimately worse.[19] But again: they attempt no such argument. So, in short, the text of the *Crito* gives us no reason to ascribe to Socrates a preference for democracy over other systems of government.

---

[16] See K.J. Dover, "The Freedom of the Intellectual in Greek Society," *Talanta*, VII (1976): 24-54.

[17] See the excellent discussion in Kraut, *Socrates and the State*, pp. 224-26.

[18] See *Gorgias* 503B-D, 515C-516D.

In short, Vlastos's argument for the democratic structure of the elenchus is powerful; his external evidence concerning Lysias and Chaerophon is extremely telling; but his argument from the *Crito* is far from airtight, and one should probably prefer an account like Kraut's, which ties Socrates' choice to other features of Athenian ways of life.

V

"The *Protagoras* and the *Laches*" discusses a problem in Socrates' definitions of courage. After getting Laches to agree that courage is wise endurance (192C-D), Socrates does not agree with him, as one might expect. Asking, "Wise with regard to what?" he goes on to produce a series of cases of wise endurance that we would not call courage: (1) a financier wisely perseveres in putting his money into something that he knows will be a good investment; (2) a doctor wisely persists in maintaining the regimen required for the patient's illness; (3) a military commander wisely persists in attacking an opposing force that he knows to be weaker than his own; (4) a cavalryman, skilled in horsemanship, persists in doing battle against an opponent unskilled in that craft; (5) a peltast, similarly; (6) an expert swimmer dives into a cistern more steadfastly than the nonexpert.

What is Socrates getting at here? One obvious answer springs to mind. None of these cases involves risk or fear. The most natural way of reading the passage is that Socrates wants to add to the definition the fact that wisdom must concern "fearful and confidence-sustaining things," "expected evils" (added, in fact, in the definition at 199A). This obvious answer is not even mentioned by Vlastos, who goes directly after something far more subtle. He concludes that the six cases are all cases of technical knowledge, whereas courage, for Socrates, concerns moral knowledge.

Here, as often, Vlastos assumes that it is clear what the distinction between the moral and the nonmoral is, and that the Greeks had such a clear distinction. This seems to me an unwise assumption. But let us suppose that what Vlastos means by 'moral knowledge' is "knowledge of the ends of human life," and that by 'technical knowledge' he means "knowledge of instrumental means to these ends": it is not clear that this really helps us see what Socrates is driving at if we do not mention risk. The reason that courage requires knowledge of ends is that, unlike the examples of prudent and risk-free decision cited by Socrates, it does involve putting one's safety on the line, and this naturally raises the question: When is it worth doing that, and for what? It is for that reason that it seems plausible to say that true courage requires knowledge of ends: if one risks one's safety without insight into what is worthwhile and what is not, it is plausible to say that this is not a genuinely virtuous action.

---

[19] This is the solution favored by Kraut.

Vlastos now turns to a detailed analysis of the final argument in the *Laches*, in which Socrates scrutinizes the definition of courage as "knowledge of fearful and confidence-sustaining things." Scholars have long been divided about what premise of this argument is the intended refutand. Terence Irwin, Christopher Taylor, and Terry Penner have argued that it is the second premise: "Courage is a part of virtue." Vlastos alludes to his earlier disagreements with these scholars over the understanding of the "unity of virtue" doctrine—which Vlastos takes to assert only a mutual biconditional implication among the virtues, his opponents to assert a more thoroughgoing unity in which the various virtue terms are different names for one and the same thing. Those who were not convinced by Vlastos's earlier arguments will find little here that might change their view. The main new point Vlastos stresses is the apparently serious use of a multipart account of the virtues in the *Meno*, which both Vlastos and Irwin agree to be later than the *Laches*. To Irwin's claim that Socrates does not actually advance this view, but simply allows it to Meno in order to refute him, Vlastos says that this would be to ascribe to Socrates a kind of "make believe" inconsistent with his elenctic sincerity. Here, we have another example of the interdependence between Vlastos's two books: the debate cannot, then, be adjudicated without full consideration of Vlastos's earlier position on Socratic irony and his treatment of the interlocutor. To my mind, Vlastos's treatment of the *Laches* was well worth publishing as the record of work in progress, but it is far from being a finished or completely convincing argument.

VI

"Socrates and Vietnam" is an address given by Vlastos to graduate degree candidates at Berkeley in May 1987. He begins with the *Phaedo*'s famous description:

> Such, Echecrates, was the end of our companion—a man, we might say, who of all those we came to know was the best and, in any case, the wisest and the most just (118A).

Was Socrates really this good, Vlastos asks? We might agree that he had several recognized virtues to an outstanding degree: courage, moderation, piety. But justice? Plato's confident claim forces us to confront the fact that Socrates stayed out of politics during a key period in Athenian history. He admits that he has never advised the state—despite the fact that the Athenian democracy gave him ample opportunity to do so. During the debate prior to the disastrous Sicilian Expedition and during the morally fraught deliberations about the treatment of the citizens of Mytilene, never once did this allegedly just man come forward to give the public the benefit of his deliberation and insight. "You know," he says in the *Apology*, "that if I had tried to do politics long ago I would have perished long ago and done no good to you or to myself" (31D). As Vlastos says: "the defeatism of this retrospective judgment is unwarranted"(129). The evidence was that a strong and rational voice could be heard (not least in the Mytilenian debate, where the assembly was in fact persuaded to reverse its brutal decision).

Socrates' decision, Vlastos argues, can be understood by thinking about his conception of his philosophical mission. "Philosophizing," in Socrates' view, requires addressing people one by one; public speaking would have been at best a distraction from this all-important mission. (Vlastos might here profitably have discussed the careers of later admirers of Socrates, such as Cicero, Seneca, and Marcus Aurelius, who drew very different conclusions about the merits of public service.[29])

Socrates' choice, Vlastos now argues, is similar to that faced by many scholars in the modern research university: their pursuit of truth is a vocation that demands all the time they can get for it, and that is never enough, given the necessary demands of teaching and university service. (Vlastos might have added that it is much more difficult for them than it would have been for Socrates to find occasions to address their society and to shape public policy.) Many scholars understandably follow Socrates' course, remaining silent on important public issues.

Vlastos now turns to his own history. Early in his career, in Canada, he became deeply involved in political work; his research was crowded out. When he moved to the United States, during the McCarthy era, he was questioned by the Federal Bureau of Investigation in a way that made it clear to him that political engagement while an alien here would be unwise. "So...energies previously siphoned off in nonscholarly speaking and writing could now be kept for research in classical philosophy. My scholarship prospered" (132). Unfortunately, Vlastos never considers ways in which scholarship itself may contribute to social justice, something about which many philosophers in the tradition, from Plato and Aristotle to Rawls, have spoken, defending the idea that even an abstract piece of philosophical work may in some instances provide a valuable public service.[21]

---

[29] See the good historical discussion in Miriam Griffin, "Philosophy, Politics, and Politicians at Rome," in J. Barnes and M. Griffin, eds., *Philosophia Togata: Essays on Philosophy and Roman Society* (New York: Oxford, 1989), pp. 1-37; and my "Lawyer for Humanity: Theory and Practice in Ancient Political Thought," *Nomos*, XXXVII (1995): 181-215.

His "cop-out from dissent" ended, Vlastos says, with Vietnam.[22] "Seeing my students and my own son required by law to become uniformed killers in a cause which they despised put an end to my qualms" (132). Despite some degree of personal risk, Vlastos agreed to move the anti-Vietnam resolution of the American Philosophical Association in December 1966, when, as then President of the Eastern Division, he was the person best placed to do so. He expresses his similar stance of protest with regard to United States activities in Central America.

Vlastos's advice to the students is "to accept what Socrates never did: the concurrent, not easily harmonized, claims on us of the intellectual's lonely search for truth and the corporate struggle for justice" (133). Both claims are morally inescapable. Citing Abraham Lincoln, he concludes that "to sin by silence when they should protest makes cowards of men" (133). Socrates, he says, would have been both wiser and more just, had that truth "enlightened his moral vision."

This address provides a most appropriate conclusion to a career dedicated to both wisdom and justice.

MARTHA C. NUSSBAUM

University of Chicago

---

[21] See Rawls, *Political Liberalism* (New York: Columbia, revised paper edition 1996), p. lxii.

[22] Vlastos simplifies here, omitting his own courageous role in opposing McCarthyism at Princeton during the 1950s.

BOOK REVIEWS

The Struggle for Recognition: The Moral Grammar of Social Conflicts. AXEL HONNETH. Cambridge: Polity Press, 1995. xxi + 215 p. Cloth $62.96[*]

Axel Honneth is Professor of Philosophy at the University of Frankfurt, a central figure on the European philosophical scene, and perhaps the most prominent member of the third generation of figures associated with the so-called Frankfurt school. Honneth's book, *The Struggle for Recognition: The Moral Grammar of Social Conflicts* was originally published in German in 1992.[1] Taking its inspiration from G. W. F. Hegel's idea of the struggle for recognition, it presents Honneth's distinctive version of critical theory (the approach to the study of society developed by the Frankfurt school), and promises a bold vision of social change, human history, and the good life.

I

Honneth wants to provide the outline of a social theory of a particular kind. His aim is to supply the foundations for what he calls (1) a "social theory with normative content" and (2) a "critical social theory." Honneth's project is thus explanatory, normative, and critical; each of these features being closely interlinked. The project's basic explanatory aim is to account for the "grammar" or "logic," which is to say, the *structure*, of social change (Honneth uses the Wittgensteinian and Hegelian terms interchangeably; 2, 70, 144, 160, 165, 190fn. 3). Its broad explanatory aim is to account for "the moral development of human society" (93). The project is normative inasmuch as it seeks to provide a framework for the evaluation of social movements (for example, as progressive or reactionary, emancipatory or oppressive) and a conception of the good life. It is critical inasmuch as it shares critical theory's traditional normative aim of criticizing social domination.

One distinguishing characteristic of critical theory is its commitment (a legacy of Hegel and Karl Marx) to the unification of criticism and comprehension. A critical theory characteristically seeks to locate its standards of criticism "within" the structures of the society it investigates. Honneth's particular version of critical theory is dis-

---

[*] I would like to thank John Carriero, Mary Devereaux, Raymond Geuss, and Wayne Martin for their helpful comments.

[1] *Kampf um Anerkennung: Zur moralischen Grammatik sozialer Konflikte* (Frankfurt am Main: Suhrkamp, 1992).

0022-362X/97/9401/46–54

tinguished by the framing notion of *its* internal critique: recognition and the struggle for recognition (*Anerkennung* and *Kampf um Anerkennung*). On his view, "processes of societal change are to be explained with reference to the normative claims that are structurally inherent in relations of mutual recognition" (2). The struggle for recognition, he argues, provides "the actual source of motivation for social progress" (144) and "the moral force within lived social reality that is responsible for development and progress" (143). Honneth's approach to critical theory is further distinguished by his break with critical theory's traditional focus on ("concrete," "material") *interests*. Taking recognition to be a specifically *moral* notion,[2] Honneth seeks to identify a set of moral motivations for social change, to show that society has a "moral infrastructure" and that social change has a "moral grammar" (*moralische infrastruktur, moralische Grammatik*).

II

Honneth's route to his version of critical theory goes something like this: (i) contrary to the "Machiavellian-Hobbesian" tradition, social change has a moral grammar; (ii) this grammar is determined by the *struggle for recognition*; (iii) the German idealist, Hegel, offers the most sophisticated treatment of this idea in the "classical" philosophical tradition;[3] (iv) but, before Hegel's conception of the struggle for recognition can be theoretically employed, it must be philosophically reconstructed; (v) and the best way to do that is to turn to the pragmatist, George Herbert Mead. The resulting social theory will bring Hegel "up to date" (*aktualiserien*) by providing a "naturalized" or "materialist" reconstruction of his account of the struggle for recognition. It will show that there are three central patterns of recognition in modern social life: love, rights, and solidarity (*Liebe, Recht, Solidarität*).

Considered in broad outline, Honneth's approach is very appealing. The goal of constructing an alternative to the "Machiavellian-Hobbesian" conception of society as an egoistic struggle for self-preservation is familiar and attractive. Hegel's conception of the struggle for recognition offers the exciting prospect of a unitary philosophical account of the origin of normativity, self-consciousness, and society. The idea that Hegel's conception stands in need of reconstruction is uncontroversial. What about the details?

---

[2] Honneth discusses his conception of morality as based on the notion of recognition in "Integrity and Disrespect," in Charles W. Wright, ed., *The Fragmented World of the Social* (Albany: SUNY, 1995), pp. 247–60.

[3] As Honneth recognizes, Hegel's treatment of the idea of recognition (*Anerkennung*) was strongly influenced by J.G. Fichte, and the modern discussion of the theme can be traced back to Jean-Jacques Rousseau's exploration of the idea (*s'apprécier mutuellement*).

Honneth's appropriation of Hegel raises two sorts of issues: (1) the part of Hegel's corpus that provides the most appropriate starting place for a reconstruction of his conception of the struggle for recognition; and (2) the rationale for that reconstruction. Honneth's account draws on the writings of the "early Jena Hegel," the texts Hegel wrote in Jena *before* the 1807 *Phenomenology of Spirit* (*The System of Ethical Life* and the so-called *Realphilosophie* in particular). This is noteworthy because the 1807 *Phenomenology* offers Hegel's most famous explicit discussion of the topic of recognition and because the 1820 *Philosophy of Right* offers what some would argue is his most developed account of the institutions required for a stable structure of recognition. There is nothing idiosyncratic, however, about Honneth's choice of texts. Hegel's early Jena treatment of the struggle for recognition has become something of a cottage industry in recent German scholarship.[4] But Honneth's suggestion that the early Jena writings constitute *the only* legitimate starting place (5, 10, 30, 63) is striking, nonetheless. It also engages a number of standing interpretive controversies.

Honneth contends that "it is only in the years that he spent in Jena as a young philosophy lecturer" that Hegel developed "an approach whose inner principle pointed beyond the institutional horizon of his day and stood in a critical relationship to the established form of political rule" (5). The underlying idea here is, very roughly, that the early Hegel is "good" (because "critical") and that the later Hegel is "bad" (because conservative). Now, it is certainly true that the 1820 *Philosophy of Right* does not provide the sort of overt, radical repudiation of modern society to which Marx or the early members of the Frankfurt school were committed, but the claim that it does not stand in a "critical relationship to the established forms of political rule" runs against the grain of mainstream Hegel scholarship of the last forty years. Nor does Honneth take into account the case for the widely accepted image of the later Hegel as a "centerist-reformist."[5] Hegel's affirmation of "the actual" (*das Wirkliche*) never amounted to an endorsement of *existing* social practices as such. On the contrary, the *Philosophy of Right's* account of the rational structures of modern social life provides the critical framework for the assessment of "established form of political rule" Honneth claims is missing from the later Hegel. That we cannot wholly embrace this framework, accept Hegel's conception of "the actual," or buy into his view that the modern social world is worthy of reconciliation goes without saying. But the later Hegel just is not the accomodationist Honneth seems to take him to be.[6]

---

[4] See, for example, Ludwig Siep, *Anerkennung als Prinzip der praktischen Philosophie: Untersuchungen zu Hegel's Jenaer Philosophie des Geistes* (Freiburg: Alber, 1979); and Andreas Wildt, *Autonomie und Anerkennung: Hegel's Moralitätskritik im Lichte seiner Fichte-Rezeption* (Stuttgart: Klett-Cotta, 1982).

[5] For a brief excellent discussion of this aspect of Hegel scholarship, see the "Editor's Introduction" to Hegel, *Elements of the Philosophy of Right*, Allen Wood, ed., H.B. Nisbet, trans. (New York: Cambridge, 1994), pp. vii-ix.

Nor is it obvious that Hegel ever abandoned his "intersubjectivist concept of human identity" or the insight that relations of recognition are pluraform (cf. 58-62, especially 62).[7] As I read it, the *Philosophy of Right* provides an account of the institutional structures required for rational intersubjective identity formation and an argument to the effect that: (a) the rational structure of the modern family is such as to provide a framework within which agents can find emotional recognition of their psychological particularity; (b) the rational structure of civil society is such as to provide a framework that makes it possible for agents to find recognition of their legal personhood, their particular talents, skills, and achievements, and their associative affiliations; and (c) the rational structure of the modern political state is such as to provide a framework that enables agents to recognize one another as members of a politically organized community.[8] Thus, the *Philosophy of Right* includes the three patterns of recognition Honneth, quite sensibly, wants for his social theory: love (within the family); legal recognition (within civil society); and solidarity, the form of esteem that allows individuals to affirm their concrete traits and abilities and feel group pride (within civil society and the state). Honneth's decision to focus on Hegel's early Jena writings is, of course, perfectly legitimate, but he has not established that these writings contain an insight that cannot be found in the later texts.

The basic idea that motivates Honneth's reconstruction—that it is essential to extract an account of *human* recognition from Hegel's general conception of *Geist* (spirit, mind)—is perfectly reasonable. But some of the specific reasons Honneth adduces for his reconstruction require close examination. He claims, for example, that Hegel did not view "intersubjective relationships as empirical events within the social world" (68). It is certainly true that, for Hegel, recognition is part of the philosophy of spirit and as such is not *merely* empirical. It is also true that Hegel did not present his account of the development of spirit as empirical history or social theory. But it does not follow from this that Hegel did not think that relations of recognition took place between human beings as empirical events. Why Honneth attributes this bizarre view to Hegel is puzzling. The "idealism" Honneth wants to avoid *is* an idealism to be avoided, but it is not a form of idealism Hegel ever held.

---

[6] If Honneth's view is that the later Hegel is normative but not critical (where being 'critical' means something more than "actively denouncing one's society"), we need to be told more about what Honneth's conception of "the critical" is.

[7] Here Honneth follows the Theunissen-Hösle view, according to which Hegel's philosophical development is marked by a tendency to purge ethical life of all forms of intersubjectivity, replacing reciprocal relations among individual human beings by the "monological" self-relation of *Geist*. See Michael Theunissen, "Die verdrängte Intersubjektivität in Hegels Philosophie des Rechts," in Dieter Henrich and Rolf-Peter Horstmann, eds., *Hegels Philosophie des Rechts* (Stuttgart: Klett-Cotta, 1982); and Vittorio Hösle, *Hegels System, Volume II: Philosophie der Natur und des Geistes* (Hamburg: Meiner, 1987).

[8] See my *Hegel's Social Philosophy: The Project of Reconciliation* (New York: Cambridge, 1994), pp. 196-97, 200, 253-54.

No less troubling is Honneth's suggestion that Hegel has no social psychology. This, I take it, is the force of his claim that Hegel's approach is bound by a "commitment to the teleological framework of a developmental theory that allows the ontogenetic process of identity-formation to flow directly into the formation of social structures" (69) and his suggestion that it is only by turning to someone other than Hegel that we can identify the "psychological mechanism" (75) that explains the development of self-consciousness. But Hegel's account of the struggle for recognition and its role in the formation of self-consciousness is *itself*, among other things, a piece of social psychology—a highly abstract one to be sure, but a piece of social psychology, nonetheless.[9]

Now, in other places, Honneth seems to grant that Hegel *has* a social psychology, but also to maintain that it is insufficiently empirical and scientific for the purposes of social theory (75, 80, 92). So Honneth's fundamental *philosophical* objection is to the *kind* of social psychology that Hegel provides. This, in fact, is the main consideration that leads him to Mead. Honneth's claim is that Mead's social psychology, which holds that "the mechanism of personality formation consists in one's learning to conceive of oneself from the normative perspective of one's neighbour" (77), provides an empirical and scientific reformulation of the Hegelian insight that "human subjects owe their identity to intersubjective recognition" (71). But this leads to the question: Is Hegel being reconstructed or replaced? A further, absolutely central question is: What conditions must a theory satisfy in order to be counted as empirical? And: What does Honneth mean by 'empirical'? What makes Mead empirical and Hegel not? Without an answer to these questions, Honneth's position is difficult to assess.

---

[9] The claim that Hegel had a social psychology is, of course, anachronistic, since the discipline of social psychology did not exist in Hegel's time. In saying that Hegel has a social psychology, we impose our contemporary philosophical vocabulary, which is to say, we "reconstruct" his thought.

The most important philosophical question concerning Mead has to do with *normativity*. Hegel maintains that the structure of self-consciousness contains an element that (in some way) commits agents to mutual recognition—to recognizing the practical claims of others as having equal standing with their own. What precisely this claim means and how exactly the argument is supposed to work are complicated matters, but it is clear that Hegel thinks that there is some kind of "logical" (rational, philosophical) connection between the structures of self-consciousness and the commitment to mutual recognition. This appears to be one of the features of Hegel's conception of recognition that Honneth wants to preserve (37, 44, 46, 92). Indeed, he *must* preserve it if he is to "explain processes of social change by referring to the normative demands that are, structurally speaking, internal to the relationships of mutual recognition" (2). *Does* Mead's account preserve this feature? Perhaps, but Honneth does not establish that it does. The very reason Honneth gives for finding Mead's conception so attractive—its specifically empirical character—provides grounds for doubt. It is not clear how a psychological account of recognition that limits itself to the resources provided by the science of human nature can account for the genuine *normative* force of the demand for mutual recognition. What is striking here is that Honneth seems not to have considered the possibility that Hegel himself may have been committed to the project of articulating an approach that could be characterized as "historicized naturalism"—an approach that would avoid Kant's "transcendentalism" while preserving normativity.[10]

---

[10] The phrase 'historicized naturalism' belongs to Allen Wood, *Hegel's Ethical Thought* (New York: Cambridge, 1990), p. 33. As used here (and by Wood), the term 'naturalism' refers not to the contemporary idea of epistemological naturalism, contemporary naturalistic approaches to the philosophy of science, or to what modern moral philosophers call *ethical naturalism* (a thesis about the meaning of moral terms) but rather to the general idea of a conception of ethical theory that is based on a philosophical account of human nature. Hegel's naturalism is *historical* inasmuch as he thinks of human nature (and reason) as evolving and developing over the course of human history. For an illuminating discussion of what Wood calls *classical naturalism* (that is, the naturalism of Plato and Aristotle) and the features that differentiate it from modern ethical naturalism, see Julia Annas, *The Morality of Happiness* (New York: Oxford, 1993), pp. 135-41. For a stimulating recent attempt to construct a classically-inspired naturalism that would overcome the Kantian dualism between "scheme and content" while preserving normativity, see John McDowell, *Mind and World* (Cambridge: Harvard, 1994). McDowell explicitly acknowledges the affinities between his project and Hegel's (p. 11). For an informative discussion of the development of scientifically-inspired "post-Fregean" naturalism and the problem of normativity, see Philip Kitcher's "The Naturalist's Return," *The Philosophical Review*, CI, 1 (January 1992): 53-114.

III

The key to Honneth's analysis of social conflict is provided by his typology of the forms of recognition and the corresponding forms of "disrespect" (*Mißachtung*). They supply the "dynamic element" (92) of Honneth's explanatory story. The two historically significant forms of recognition are legal rights and solidarity; the two historically significant forms of disrespect are the denial of rights and denigration (133). The basic idea is remarkably simple. When individuals are "structurally excluded" (163) from the possession of legal rights or when their ways of life are denigrated, *their feelings are hurt* (163, 165, 168). If they possess a vocabulary allowing them to grasp that their personal disappointments are shared by other members of their group, the "motivational impetus" (138) for a collective struggle for recognition is in place. This may, in turn, lead to social action, resulting in the expansion of relations of recognition (163, 168). The sequence of disrespect, demand for recognition, and social change constitutes the "developmental logic" (139) of collective movements. This, apparently, is what Honneth's conception of a "moral logic" and "moral grammar" comes to. The underlying idea here—that moral feelings *can* motivate social struggles—is plausible enough, especially since Honneth emphasizes that the experience of disrespect does not *always* lead to resistance (138), but it does not seem to have much bite. Nor is it obvious that we need Hegel or Mead to see that feelings of moral outrage can motivate social action.

In a rather surprising turn, made in the penultimate chapter of his book, Honneth explains that his model of social change is meant to "extend" rather than "replace" the more familiar model that explains social change in terms of the pursuit of interests (165). This crucial qualification sits rather uneasily with Honneth's ambitious use of the definite article in connection with "grammar of social conflicts" in the book's subtitle. *The* moral grammar of social conflicts appears to have become *a* grammar. Now, it may be that Honneth thinks there is one *moral* grammar of social conflicts—in addition to a distinct *nonmoral* grammar of social conflicts. But he neither develops nor defends this point. If recognition provides one "foothold of morality within society" (138), why should there not be others? Nor does Honneth say anything systematic about the relation between the moral grammar and the nonmoral grammar of social conflict. What he says instead is that "it will be always be an empirical question as to the extent to which a social conflict follows the logic of the pursuit of interests or the logic of the formation of moral reactions"

(165). Which leads one to wonder what exactly Honneth's *theory* of social conflict is.

A related question is posed by Honneth's main historical thesis, "that the moral force within lived social reality that is responsible for development and progress is a struggle for recognition" (143). How exactly is this claim—which Honneth himself characterizes as "strong" (1)—to be understood? What precisely is the scope of "development and progress"? Are *all* instances of development and progress to be explained by reference to the struggle for recognition? Some? Many? And how is this account of the genesis of development and progress to be reconciled with the view that social conflict is characterized by a "logic of the pursuit of interests" as well as a "logic of the formation of moral reactions"? It is clear that Honneth wants a general conception of history (92, 93, 168, 171), but it is not clear what that conception is.[11]

The apex of Honneth's theory, his account of the good life, also gives rise to a fundamental question. Honneth calls this account "a formal conception of ethical life." It provides the normative standard for the assessment of social struggles and the "hypothetical endpoint" toward which the history of social struggles is to be viewed as moving (171). Honneth first defines "ethical life" (*Sittlichkeit*) as the totality of intersubjective conditions necessary for individual self-realization (*individuellen Selbstverwirklichung*) (173), understood as "the process of realizing, without coercion, one's self-chosen life goals" (174). He then identifies these conditions with the three patterns of recognition constituted by love, rights, and solidarity (174). Thus, for Honneth, love, rights, and solidarity are the intersubjective conditions of individual self-realization. Honneth's conception of ethical life is supposed to be "formal" in that it consists of "structural elements" that can be "normatively extracted from the plurality of all particular forms of life" (172). The three patterns of recognition are held to be defined "in a sufficiently abstract, formal manner to avoid raising the suspicion that they embody particular visions of the good life" (174).

Honneth's conception of the good life is appealing. It might be thought of, very roughly, as a combination of the J. S. Mill of *On Lib-*

---

[11] In the sentence immediately following his statement of his main historical thesis, Honneth suggests that he does not have (or want) a philosophy of history (143). What he presumably means by this, however, is that he does not have a theory of history which holds that there is a fixed goal of human history or predetermined final state—*not* that he does not have a general conception of the development of society, which is to say, of history.

erty and the Marx of the review of Mill's *Elemens d'économie politique*. Another question is how formal it is. Honneth's strong individualism sure looks like a particular vision of the good life. His "formal" conception of ethical life appears to be just the sort of conception John Rawls[11] regards as comprehensive. So it certainly raises *the suspicion* that it embodies a particular vision of the good life. But the basic question is this: What is doing the normative work in Honneth's critical theory? At the outset of the discussion, it appeared that *recognition* was the basic normative notion of the theory. But, then, when we get to the final chapter of the book, it seems that *self-realization* is the basic normative notion, and that the normative significance of recognition is derivative, recognition being important as only as a precondition of self-realization. So it is not completely clear what Honneth's normative standpoint is.

Given Honneth's stated aims and the ongoing nature of his project, it would be unfair to expect a fully developed account. But, at this point, many of the key components of his critical theory remain to be worked out.

MICHAEL O. HARDIMON

University of California/San Diego

---

[11] *Political Liberalism* (New York: Columbia, 1993), and "Reply to Habermas," this JOURNAL, XCII, 3 (March 1995): 132-80.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 2, FEBRUARY 1997

ANALOG REPRESENTATION BEYOND MENTAL IMAGERY *

Among the many ideas that computer science has contributed to a general theory of cognition, the distinction between analog and digital representation is one of the earliest and most fundamental. There has been a shift, however, in the way this distinction has been taken, largely because of its importance in the debate on the nature of mental imagery—a shift which may have lost a dimension of the analog that would have been valuable in resolving the imagery debate. The later view, which I shall refer to as the *sense interpretation* of the analog, has relied on the intuition that the analog has some essential connection to sensory cognition (or to senselike cognition, such as imagination). Thus, *continuity* and *density* (the latter is Nelson Goodman's term), which are characteristics of the spatiotemporal media of sense perception, have been taken to be necessary properties of the analog.

The original view, however, which I shall call the *model interpretation*, relied more on the notion of maps and models as prototypically analog forms of representation. Thus, in contrasting a digital and so-called analog computer with one another, it was commonplace to take the latter as a reproduction of what it was designed to represent. A wind tunnel that reproduced the conditions of actual aerodynamic flight, coupled with scaled-down models of aircraft, is an analog computer where results could be attained without "digital computation." A slide rule was analog (whereas an abacus was digital) because it modeled the serial arrangements of number series, instead of relying on calculational algorithms which themselves required abstract, conceptual representations of certain properties of numbers.

An analog computer could model the relations among mathematical variables by reproducing the actual variation of those variables in

0022-362X/97/9402/55–84

another medium (for example, voltages, or purely mechanical functions); that is, such a "computer" physically instantiated the *values* of those variables. In a digital computer (and in digital representation), on the other hand, it was possible to represent the variables without instantiating their values. To be able to manipulate "types" without having to deal with any "tokens" is essential both to digital calculation and to intentionality.

It was the difficulty inherent in characterizing intentionality and the cognitive status of abstract "types," I believe, that led to the superficiality of the sense interpretation, which relied on what it took to be a basic difference in signs or symbolic media as a sufficient basis for distinguishing analog from digital: digital representation came to be taken as necessarily discontinuous or discrete, while the analog was taken as necessarily continuous. I shall argue that the discreteness of digital symbols is necessary because they stand for disjoint *variables*, as opposed to standing in for sets of *values* of variables; that is, it was the variable/value distinction (on which the model interpretation rests) that was the more principled ground for this basic distinction between forms of representation.

It was natural, I suppose, that analog "symbols" should have been taken to be necessarily continuous, for it was natural to take these sets of values as necessarily physically instantiated, that is, as spatiotemporal. But this was an error; for there are obviously countless sets of values whose variables do not vary continuously (the series of whole numbers, for example). If the variable/value (or type/token) distinction is the basis for the digital/analog distinction (as I believe it to be), then it is the existence of discontinuous variation that should be sufficient to disqualify continuity as a necessary property of the analog and should count as negative evidence for the sense interpretation.

The variable/value distinction (with discontinuous as well as continuous variation) is explicit, however, in the model interpretation. An analog map or model does not provide an algorithm, rule, or variable to represent the concrete instances that it is designed to map or model, but rather reproduces some aspect of the relations among those instances/values. A two-dimensional drawing is still a reproduction: there is some *identity of relation* (relations among the instantiated values) between the original and the reproduction, if not an identity of relata. It is this relational identity (which entails neither continuity nor sensible content), I want to claim, which is sufficient to constitute a map or model—and sufficient for "analog representation" as well. Algorithms, rules, or variables, on the other hand, are not reproductions of (some aspects of) instantiated values and do not rely on such relational identity.

I want to review the general outline of the imagery debate and examine some traditional conceptions of the analog in order to show how the differences between the two general interpretations I have outlined have influenced the discussion. If we understand the analog in the way I have proposed, even purely abstract conceptual analogies can be taken as instances of analog cognition; this would disengage the question of the nature of the analog from the debate on mental imagery and would have important consequences for a general theory of representation. So, even if Stephen Kosslyn[1] is correct in declaring the imagery debate "resolved," this would not mean (as he suggests) that we have settled the debate regarding the basic differences between analog and digital forms of representation.

I. MENTAL IMAGERY AND THE PROBLEM OF ANALOG "REPRESENTATION"

Traditional cognitive theory holds that knowledge processes are analyzable as formal operations on symbolic representations. Jerry Fodor and Z. Pylyshyn's[2] representational theory of the mind takes representation, strictly speaking, to be digital ("symbolic," "encoded") in nature. Analog representation, on this view, is impossible.[3] They argue that so-called "analog" properties are really reducible either to (a) the physical or architectural properties of the sensory detection systems of the organism, which constrain representational processes without constituting them, or to (b) digital representations, or to a combination of the two. Digital representation, Pylyshyn argues, is "cognitively penetrable":

> If we can set up situations demonstrating that certain stimulus-response regularities can be altered in ways that follow...rational principles, we say that the input-output function in question is *cognitively penetrable*, concluding that at least some part of this function cannot be explained directly in terms of properties of the functional architecture; that is, it is not "wired-in" but requires a cognitive, or computational, or representation-governed, explanation (*op. cit.*, p. xvii).

An analog process, on the other hand, "is one whose behavior must be characterized in terms of lawful relations among properties of a particular physical instantiation of a process, rather than in terms of rules or representations (or algorithms)."[4] A psychological transducer which functions as a "bridge from physical to symbolic" is, like other analog-to-digital converters, not itself symbolic because it is part of the functional architecture.[5] Perception is the "product of both transduction and cognition" and must include inferential, propositional, and/or computational components.[6]

---

[1] *Image and Brain: The Resolution of the Imagery Debate* (Cambridge: MIT, 1994).

[2] See Fodor, "Methodological Solipsism Considered as a Research Strategy in Psychology," *The Behavioral and Brain Sciences*, III (1980): 63-109; Fodor and Pylyshyn, "How Direct is Visual Perception?" *Cognition*, IX (1981): 139-96; and Pylyshyn, *Computation and Cognition* (Cambridge: MIT, 1984).

[3] This helps explain why the term 'analog' is frequently used in these discussions, but the term 'digital' hardly ever.

But how do we deal with the apparently irreducible spatiotemporal qualities of analog images (and the seemingly physical principles governing their scanning, rotation, and so on)? Unless mental images can be reduced to physical and/or representational form, Pylyshyn claims, they are "intellectual orphans" and serve to explain nothing (*op. cit.*, p. 51). Images cannot be true representations, he argues, because representational symbols are atomic and discrete while images have continuous properties. Further, he asks, how could we possibly compute over "analog" symbols? There would be nothing equivalent to a "calculus" of such symbols (*op. cit.*, pp. 159, 199). Many experiments that claim to show that processing via images is different from processing by verbal descriptions, Pylyshyn claims, fail to take into account the possibility that subjects were replicating (through a sequence of steps in time) what they would do if they were actually perceiving, instead of actually using an imaginal medium which constrains their processing (*op. cit.*, p. 127). In all of this, Pylyshyn appears to subscribe to the sense interpretation of the analog; but he denies that the analog, so interpreted, really has any cognitive status.

Fodor is more open to the possibility that images have an important cognitive role, but he favors treating all images as "images under descriptions" whose distinctive features (including their capacity to resemble) and representational power derive mainly from their descriptive association; he has admitted, however, that it is possible that some images are not generated from descriptions and thus never had a digital (descriptive) storage phase.[7]

---

[4] Fodor and Pylyshyn, p. 157. Ned Block, who ultimately rejects Pylyshyn's complete descriptionism, still holds a partially physicalist conception of the analog. *Primitive processors*, that is, the bottom level processors "whose properties can only be explained nomologically," are analog—"Mental Pictures and Cognitive Science," *The Philosophical Review*, XCII (1983): 499-541, here pp. 523, 533. Nomological explanations are apparently explanations in terms of purely physical processes, that is, "not explanations in terms of manipulation of representations" (p. 533).

[5] Pylyshyn, p. 154.

[6] Pylyshyn, pp. 134-35; Fodor, *The Language of Thought* (New York: Crowell, 1975), pp. 44-50; Fodor and Pylyshyn, pp. 139-40.

[7] *The Language of Thought*, p. 194.

Those who have defended the idea of genuinely analog representations, emphasizing the qualitative differences in task strategies suggested by many empirical studies, have not clearly articulated a theory supporting their view and are consequently hard pressed to meet these sorts of objections. A. Paivio[8] has defended a "dual coding theory." His basic assumption is that "there are two classes of phenomena handled cognitively by separate subsystems, one specialized for the representation and processing of information concerning nonverbal objects and events, the other specialized for dealing with language" (*ibid.*, p. 53). He also seems to adopt the sense interpretation of the analog. The two subsystems are organized differently: "component information in higher-order nonverbal units is synchronously organized (permitting parallel processing up to some informational limit), whereas verbal components are sequentially organized (implying sequential constraints on intra-unit processing)" (*ibid.*, p. 59). My mental image of my living room, Paivio explains, "is synchronously available in the sense that I can describe its contents beginning at any point and from any perspective in my mental space. This implies that the mental representation of my living room 'contains' information that is simultaneously available. It does *not* mean that all the information can be accessed or processed simultaneously" (*ibid.*, p. 60).

Paivio does not provide a detailed theoretical account of the difference between synchronous and sequential organization beyond this simple description, nor does he consider how this difference between sequential and synchronous organization might be relevant for a general theory of representation. As we have seen, he associates sequential organization with verbal processing and synchronous organization with nonverbal processing and, contrary to the claim of J. Kirby and J. Das,[9] denies that there is any hard evidence of sequential nonverbal and synchronous verbal organization, although he admits their possibility (*op. cit.*, p. 112). This is significant in that Paivio thereby at least opens the door for the model interpretation of the analog.

Kosslyn has argued repeatedly for the existence of nondigital representation in mental imagery and has responded to many of Pylyshyn's objections with more refined empirical testing. In his most recent comprehensive statement, he has declared that, given the evidence and the sort of interpretation he has proposed, "we can stop debating about the fundamentals," that is, the mental imagery debate has now been "resolved" (*op. cit.*, p. 406). He takes this view because he believes the principal objections to the possibility of nondigital representation (from Pylyshyn and others) have been dealt with. Further, even though proponents of the digital-only[10] view of representation may be able to salvage their theory by relying on various theoretical amendments to explain recent empirical data, that theory grows less simple, and can rely on no substantial experimentation that might favor it over its competitor.

---

[8] *Mental Representations* (New York: Oxford, 1986).

[9] "Comments on Paivio's Imagery Theory," *Canadian Psychological Review*, XVII (1976): 66-68.

While I agree with Kosslyn that nondigital (or nonpropositional) representation exists, I do not agree that such nonpropositional mental states are confined to sensory or spatiotemporal media, as the sense interpretation would have it; that is, I would argue that analog representation may be found in purely abstract thought as well. Can the debate on the nature of mental representation be over with this issue still on the table? Pylyshyn's objections are based on his theory of mental representation; while that theory has not, it seems, been supported by the evidence to which Kosslyn points, Pylyshyn and others have looked for ways to save it because of its theoretical plausibility. Kosslyn, on the other hand, has more evidence on his side, but does not, I believe, have an equally plausible theory of representation (which is broader than a theory of imagery). It is, to be sure, philosophers who worry about theoretical cogency more than most, but the debate cannot be resolved without such a theory in place. And we should not think that science can "move on" while philosophers continue to fret,[11] for an adequate theory will in all likelihood have interesting and fruitful empirical implications—for example, the possibility that there are identifiable mental phenomena in abstract thought in which nonspatiotemporal analog mapping is employed.

As we shall see, some aspects of Kosslyn's description of depictive representation are quite relevant to the model interpretation of the analog; but we must wait until we have isolated the essential properties of the analog before we consider them.

---

[10] Kosslyn uses the terms 'depictive' and 'propositional' rather than 'analog' and 'digital'.

[11] In what looks like an afterthought, Kosslyn qualifies his declaration of the resolution of the debate: "I fully expect philosophers to continue to debate the matter; after all, that is their business. But I believe that enough is now known to begin to build on and accumulate theory. Philosophers are still arguing about the status of basic constructs in physics...and biology..., but that has not gotten in the way of genuine progress in those fields. I do not mean to imply that philosophers have not played a useful role in the past or that they shall never do so again" (*op. cit.*, p. 409, note 1). Despite his faint praise in the last sentence, Kosslyn does seem to feel that continuing to work on the theoretical problems that still concern philosophers would be more of a hindrance than a help—as if the "business" of philosophy is debate for debate's sake.

Many of the claims regarding analog representation in the imagery debate depend on conceptions that are at best intuitively held. Let us examine the more systematic attempts to characterize the nature of the analog.

1. Resemblance. Analog representations resemble their referents; digital representations do not. Depictions are analog while descriptions are not. Full-scale three-dimensional copies and models resemble their prototypes to a greater degree than do two-dimensional photos and paintings. Caricatures and sketchy diagrams are still less like the originals. Yet resemblance is a relative and notoriously imprecise notion. A caricature of Socrates will resemble its subject just insofar as it successfully represents what the artist (and the viewer) take to be salient features (for example, a snub nose and bulging forehead with receding hair), but such a determination of salience will depend on the purpose of the representation as well as on the population of which Socrates is taken to be a member (if it were filled with snub noses and bulging foreheads, other features would have to be chosen for a caricature). A geographic map, like a caricature, selects certain dimensions for its analog representation (scaled-down representations of roads, rivers, cities, and the like) and *to this extent* resembles its referent. (Because of its additional linguistic components, of course, a map is also descriptive/digital in character.) In sketchy drawings or diagrams that an inventor might scribble on a napkin, even proportion and scale may be neglected; but we could still judge that the sketch is still analog as long as it successfully reproduces the location and position of functionally important components relative to one another.

The selection of highly esoteric bases of resemblance may give the impression that resemblance has no really objective foundation and derives completely from one's point of view; but the point of even such esoteric mapping is still to preserve a delimited identity with the original, that is, to have the original (in some respect) to work with. None of this forces us to deny that, in every analog representation, the nonidentical aspects of that representation are vital for the analog's utility; even physical models are useful precisely because, insofar as they are scaled down (and to that extent nonanalog), they are more manageable and manipulable. It is, in fact, this difference between referent and representation which may establish an analog representation as symbolic, as something that "stands in" for its referent without being a mere double of it. Yet this may not be enough to qualify the analog as genuinely representational in the sense required by Fodor and Pylyshyn.

Goodman,[12] whose account of the analog-digital distinction is the starting point for most other analyses, denies that topological diagrams are analog. Electrical diagrams and diagrams of carbohydrates are, he says, purely digital (*ibid.*, pp. 170-71). We are surprised by this, he continues, only if we cling to "some loose notion of analogy or resemblance" as the basis for the analog. I would argue, however, that the imprecision of this "loose notion" should not in itself deter us from classifying such diagrammatic resemblances as analog. The arrangement and relative position of atoms in a complex molecule, or of resistors and capacitors in a radio, may very well be depicted in a faithful "resembling" diagram without regard for the internal constitution of their individual components. That is, they may still function as maps in a legitimately analog sense, even if individual symbols are digital. Consider the Peutinger table, the unique medieval copy of an antique map of the Roman empire: one can only barely recognize the shapes of land masses as those of the Mediterranean regions; the nonsystematic distortions[13] of these shapes make it unreliable for modern use. But if our concern were simply which cities were between which other cities, it would provide analog information much in the way a wiring diagram would. An important dimension (or dimensions) of resemblance would be retained to the neglect of other irrelevant features, a selection which, to be sure, requires a discrimination of features which is the hallmark of abstract thought; but this discrimination is a constraint on the production of an analog map, and is not sufficient for its constitution.

Any spatiotemporal medium may be utilized in an analog representation for any other such medium. Multiple transformations or transductions across such media make complex communication possible. The spatial variations in the grooves cut into an (analog) phonograph record are a mapping of the temporal variations of electrical impulses that guided the groove cutter responsible for the template, and both of these are mappings of the patterns of air waves constituting the sound. The groove cutter and phonograph stylus are the transducers which translate in opposite directions, as are the microphone and the speaker. Intensive as well as extensive continua are utilizable in this way; the relation between voltage and the spatial position of a voltmeter indicator is analog, for example, as is the mapping we may artificially form between varying pressure on a stationary surface and corresponding variations of light intensity. Do the variations of light intensity "resemble" the variations in pressure? We may very well say so, as long as we recognize the extraordinarily restricted dimension along which such resemblance is to be judged.

---

[12] *Languages of Art* (Indianapolis: Hackett, 1976, 2nd ed.).

[13] Map "projections" are systematic distortions.

Consider a map of the United States in which the sizes of the states are in proportion to their respective populations or annual per capita incomes. Here, we have the barest quantitative thread of resemblance: the bigger the state, the wealthier it is. The resemblance or identity is established, not at the level of individual symbols (a given state does not resemble its wealth), but at a second-order level of differences or variations in the quantities. This distinction is identical to that proposed by Roger Shepard[14] between "first- and second-order isomorphisms," to which I shall return. We may ask at this point, however: Is continuity a necessary condition for the mapping on which such resemblance is based?

2. *Continuity.* Analog media are continuous, it is often claimed, while digital media are discrete. Even the simplest characterization of the distinction between discrete and continuous media given in introductory computer-science textbooks will strike the contrast between the functions of switches and tuners, the former exhibiting a binary "on-off" function with no intermediate states and the latter permitting an indefinite continuous selection of states. Analog-to-digital conversion, requiring the translation of continuous media into discrete states, will, according to this view, require the deletion of some material (the introduction of gaps) in the representation. Unfortunately, many discussions of such conversion give the impression that the sound reproduced from a digital recording, for example, might be taken as a digital representation simply because of the fact that, unlike the original sound, it has been segmented into discrete units. Those who take television or newspaper pictures to be digital because of the discreteness of the lines and dots into which the picture is divided (and not because of the process of producing these discrete elements) make the same mistake.[15] Reproductions with gaps are not as comprehensive as reproductions without them; but they are both analog; that is, they are both reproductions. This is not to deny, however, that there is some connection between discreteness and the digital, as we shall see in the following section.

---

[14] "Form, Formation, and Transformation of Internal Representations," in R. Solso, ed., *Information Processing and Cognition: The Loyola Symposium* (Hillsdale, NJ: Erlbaum, 1975), pp. 87-122.

[15] For example, Elliott Sober, "Mental Representations," *Synthese*, XXXIII (1976): 101-48, here p. 141.

Another reason that analog representations are sometimes mistaken as digital is that analog representations may retain the form but not the material of what is represented (a pattern of electrical impulses modeling a pattern of sounds, for example); they are thereby taken to "encode" the original, and encoding (even in this broad sense) is often taken as digital. Even if we use 'encode' in this liberal way, we would still have to admit that the sorts of encodings which merely model (reproduce) the original are qualitatively different from those which are linguistic or mathematical, and it is this qualitative distinction which is vital for a general theory of representation. I would prefer to reserve the word 'encode' solely for the latter (digital) type, but these distinctions will be clarified as I proceed.

3. Density. A digital music recording is digital, not because its sound contains gaps not present in the original: such a result could have been produced simply by taking an analog recording and introducing artificial gaps into it by selective erasures. Even if discreteness of states is necessary for digital representation, it is not sufficient.

Goodman's analysis of the analog and digital in terms of "dense" and "differentiated" notational systems goes beyond any facile characterization in terms of discontinuity and discreteness alone and provides a clearer link between the digital and symbolic (or linguistic) representation. According to Goodman, a system "is analog if syntactically and semantically dense" (*op. cit.*, p. 160), and a scheme "is syntactically dense if it provides for infinitely many characters so ordered that between each two there is a third" (*op. cit.*, p. 136). This does not in itself require, Goodman recognizes, that the scheme be continuous; for example, when "the characters correspond to all rational numbers that are either less than 1 or not less than 2," there will be gaps, and "the insertion of a character corresponding to 1 will destroy density." A dense scheme will always be *undifferentiated*, however. A scheme is differentiated (or finitely differentiated) if "for every two characters *K* and *K'* and every mark *m* that does not actually belong to both, determination either that *m* does not belong to *K* or that *m* does not belong to *K'* is theoretically possible" (*op. cit.*, p. 135-36). The characters of a differentiated scheme are therefore discrete, distinct, and theoretically identifiable. Even a scheme of two characters may still be undifferentiated; for example, in the case where "all marks not longer than one inch belong to one character, and all longer marks belong to the other." Here, it is possible that one may not be able to determine to which character a given mark belongs. A digital system will be discrete (discontinuous), but must also be syntactically and semantically differentiated.

The discreteness implicit in Goodman's notion of differentiation is tailored for service in notational systems.[16] Syntactically discrete and disjoint symbols are needed to represent that which is conceptually distinct and disjoint. It is in this sense which we should take Pylyshyn's claim that "the notion of a discrete atomic symbol is the basis of all formal understanding. Indeed, it is the basis of all systems of thought, expression or calculation for which a *notation* is available" (*op. cit.*, p. 51).[17] The word 'chair' expresses the discreteness, not of an individual chair, but of a *type* that is by nature distinct and disjoint. Differentiated linguistic symbols (composed of differentiated characters) are like conceptual switches—binary devices whereby we either classify a given object as a token of a type or do not. The "on-off" character of the switches of digital computers (or the nervous system) is significant not for their discreteness alone, but because "off" can represent "zero," "nothing," "absence."[18] Zero, because it serves to define a "place" in digital notation, may really be taken, not as a token of the type "number," but as the absence of all tokens, and thus it implicitly defines the type. That is, the conceivability of any concept depends on the conceivability of the absence of *any* individual instances or tokens, leaving only the type.

The possibility of the abstraction of types or concepts is thus often connected with a species of negation. Fred Dretske,[19] for example, describes all analog-to-digital conversion as requiring some *loss* of information:

---

[16] For Goodman, a digital system is fully notational only if it is also unambiguous and syntactically and semantically disjoint (*op. cit.*, pp. 148-54).

[17] Pylyshyn goes on to claim that, as a result, there can be little basis in the operation of thought for representation with continuous or analog properties.

[18] "A *positive procedure*" (in reading and writing), John Haugeland explains, "is one which can succeed absolutely and without qualification—that is, not merely to a high degree, or almost entirely, but perfectly, one hundred percent"—"Analog and Analog," in J. Biro and R. Shahan, eds., *Mind, Brain and Function* (Norman: Oklahoma UP, 1982), pp. 213-25, here p. 214. *Reliable procedures* are those "which, under suitable conditions, can be counted on to succeed virtually every time" (p. 215). Analog procedures, Haugeland continues, are always approximation procedures and are never positive (p. 221). This characterization combines the disjunctive ("on/off") aspects of digital systems with the representational function of abstract symbols. Even simple descriptions of digital machines refer to a new "logical" level of operation, distinct from physical analogues; see John von Neumann, *The Computer and the Brain* (New Haven: Yale, 1958), p. 9. Digital error is not a mistake of resemblance or depiction, but logical error (pp. 78-79). "The contrast between analogue and digital machines," Donald MacKay observes, "can be expressed by saying that whereas analogue computers use the *magnitudes* of electric currents (or other physical entities) to represent *numbers*, digital computers use the mere *presence or absence* of such physical entities to represent *numerals* or other symbols"—*Brains, Machines and Persons* (London: Collins, 1980), p. 45. As we move from 2 to 3 in a decimal system or from 0 to 1 in a binary system, MacKay continues, we acknowledge that there are "no *numerals* in between. Indeed, there is no defined meaning for 'in between' in such a case. The transition from one to the other, as we say, is 'all-or-nothing'" (p. 46).

[19] *Knowledge and the Flow of Information* (Cambridge: MIT, 1981), p. 141.

> Digital conversion is a process in which irrelevant pieces of information are pruned away and discarded. Until information has been lost, or discarded, an information-processing system has failed to treat *different* things as essentially the *same*. It has failed to classify or categorize, failed to generalize, failed to "recognize the input as being an instance (token) of a more general type" (*ibid.*, p. 141).

The loss of information required for the generation of a concept or type is, Dretske explains, intrinsic to any genuinely cognitive state, that is, to *representation*—to taking an object *as* something:

> If the information that *s* is *F* is never converted from a sensory (analog) to a cognitive (digital) form, the system in question has, perhaps, seen, heard, or smelled an *s* which is *F*, but it has not *seen that* it is *F*—does not *know* that it is *F*. The traditional idea that knowledge, belief and thought involve *concepts* while sensation (or sensory experience) does not is reflected in this coding difference. Cognitive activity is the *conceptual* mobilization of incoming information, and this conceptual treatment is fundamentally a matter of ignoring differences (as irrelevant to underlying sameness), of going from the concrete to the abstract, of passing from the particular to the general. It is, in short, a matter of making the analog-digital transformation (*ibid.*, p. 142).

In analog representation, on the other hand, there is no negation, no truth or falsity, nothing taken "as"—no logical type at all; that is, we are among individual tokens and values, not types and variables.

The advantage gained in utilizing notational systems is clear: we can manipulate them with greater accuracy and efficiency than analog models would allow. That is, we can *calculate* with them. "In digital devices," John Haugeland recognizes, "the main thing is eliminating confusion over the type of any token; and a primary motivation ('payoff') for this is the ability to keep great complexity manageable and reliable" (*op. cit.*, p. 219). But the disadvantage is also apparent: there is always the danger of neglecting some important feature of the referent in the representation. An analog model may not provide an abstract "principled" understanding of the referent, but a digital representation may not provide a sufficient access to that referent.

Many cognitive theorists have recognized (in one way or another) that digital representations must be rules.[9] As rules, they are able to provide an efficient means for accounting for what would otherwise be an unmanageable number of individual contingencies. Just as rules of formation are required by many grammarians to handle the unbounded set of possible sentences, so too, S. Ullman[21] suggests, a perceptual system that incorporates such rules "is a more likely candidate for coping with the enormously complex tasks of perception" (*ibid.*, p. 376). He compares this to the analogous case of integer addition:

---

[9] It is not only their assignment to their referent that establishes the semantics of representations, Douglas Hofstadter observes, but their capacity to capture a space of possibilities: "This way in which any situation that is perceived seems to be sur-

> We have seen how the addition of any two integers can be based on a restricted lookup table, augmented by the right-to-left processing rule and handling of carry. Is this mode of addition better than a large-scale table that lists directly the results of adding pairs of integers? The large-scale table has an advantage: it does not require intermediate steps and therefore offers simplicity and possibly speed. The indirect method offers a different advantage: employing only a restricted table, it is able to handle an unbounded set of inputs (*ibid.*, p. 375).

The "large-scale table" to which Ullman refers here is the array or matrix of instances formed by assigning all possible values to the rule's variables.

4. Seriality. It is, as we have seen, differentiation and not just discreteness which is necessary for notational systems. David Lewis[22] has argued, however, that there is "representation that is as differentiated and non-dense as any digital representation, and yet is analog rather than digital representation" (*ibid.*, p. 322); that is, the analog

---

rounded by a cluster, a halo of alternative versions of itself...seems to me to be the dead center of thinking"—"Artificial Intelligence: Subcognition as Computation," in F. Machlup and U. Mansfield, eds., *The Study of Information: Interdisciplinary Messages* (New York: Wiley, 1983), pp. 263-85, here p. 283. Formal tokens such as "I" or "hamburger," he claims, do not in themselves denote (p. 276). See also Dretske; and Fodor, *Psychosemantics* (Cambridge: MIT, 1987), ch. 4, and *A Theory of Content* (Cambridge: MIT, 1990), chs. 2-3. Shepard includes the "possible spatial transformations of an object as a basis for the representation of its shape"—"Perceptual and Analogical Bases of Cognition," in J. Mehler, E. Walker, and M. Garrett, eds., *Perspectives on Mental Representation* (Hillsdale, NJ: Erlbaum, 1982), pp. 49-67, here p. 60. Also, recall this decisive move in Immanuel Kant's *Critique of Pure Reason*, Norman Kemp Smith, trans. (London: Macmillan, 1963), A105-6: "It is only when we have produced synthetic unity in the manifold of intuition that we are in a position to say that we know the object. But this unity is impossible if the intuition cannot be generated in accordance with a rule by means of such a function of synthesis as makes the reproduction of the manifold *a priori* necessary, and render possible a concept in which it is united.... A concept is always, as regards its form, something universal which serves as a rule."

[21] "Against Direct Perception," *Behavioral and Brain Sciences*, III (1980): 373-415.

[22] "Analog and Digital," *Noûs*, V (1971): 321-27. Pylyshyn is similarly suspicious of restricting the analog to the dense or continuous (*op. cit.*, pp. 200-01).

is not to be restricted to dense systems. Although he does not explicitly make the distinctions on which I am relying, it is likely that Lewis takes this position because he adopts the model interpretation of the analog.

Lewis's first example of an analog yet nondense representation: in an analog computer, a variable resistor could be used which works on a continuous slide, but a discontinuous selector switch with a large but finite number of positions (each adding a 1-ohm resistance) could be used just as well. The switch positions are differentiated and nondense; yet the representation is still analog.[23] Or consider, Lewis suggests, a device that produces representations by means of the measured amount of some (dense, undifferentiated) fluid; if this fluid were replaced with discrete (nondense, differentiated) pellets, the device would remain analog.

Why are these devices still "analog"? Apparently, because the representations are capable of incremental adjustments in step with the components they are designed to model (even if the increments are discontinuous); that is, it is the modeling or mapping relation on which Lewis relies. This selection of seriality as the basis for the analog is a decisive step away from the sense interpretation of the analog, to which the criteria of continuity and density were still confined.

Goodman did examine a similar example that Lewis did not consider. A pressure gauge consisting of a circular clock face and pointer with no markings or divisions is, Goodman observes, both semantically and syntactically undifferentiated (and hence cannot be digital) (*op. cit.*, p. 157). But how does the situation change if the dial is subdivided by discrete markings? Is the symbol scheme now notational (and digital)? The reply, Goodman claims, depends on how we take the markings. If we take them only as aids in locating absolute (continuous) location on the face, the entire system remains semantically and syntactically undifferentiated, and the markings do not, despite appearances, constitute or contribute to the symbol scheme for the gauge. If, on the other hand, each marking locates the center of a region and the regions are both disjoint and separated by gaps, "the system will be notational [and digital] provided the ranges of pressure correlated with the fifty characters are also disjoint and separated by gaps, however small" (*op. cit.*, p. 158).

A more obvious case of such a notational system would be a counter that indicates the number of dimes dropping into a bank by pointing to Arabic numerals. If the count is given by a numeral displayed in an aperture, the system will be clearly notational and presumably digital. But suppose the indicator is a pointer on a circular clock face. Goodman's response is that, "if the counter with a graduated face is so read that there are just fifty syntactically disjoint and differentiated characters, and each is correlated with a different number of dimes, the system will be notational" (and digital) (*op. cit*, p. 159).

---

[23] Pylyshyn would agree (*op. cit.*, p. 200).

Goodman does not here acknowledge that these cases may really involve mixes of analog and digital elements. Just as a model molecule obviously possesses digital characteristics (for an individual ball is not taken as analog to its corresponding atom), as well an analog element (for the relations among balls in the model does reproduce the relations among atoms), so a circular clock face with Goodman's disjoint and differentiated (numbered) regions may, insofar as we consider each differentiated region by itself, be digital, and yet the relation of these regions to one another in the circle (serially, but not continuously) may still provide the basis for an analog representation.

Lewis may be right, therefore, that Goodman's view would fail to distinguish purely digital systems from other mixed digital/analog types. Consider a clock face which records just the twelve hours (with no minute or second hands). Three versions are imaginable: (a) one in which there is continuous movement of the hour hand through the twelve numerals; (b) one in which the hand makes discontinuous jumps from one numeral to the next, arranged serially "clockwise"; and (c) one which is like (b), except that the numerals are arranged nonserially, for example (moving clockwise): 4, 1, 10, 8, 3, 2, 12, 5, 6, 11, 9, 7. The hand in (c) would move discontinuously from 1 to 2 to 3, and so on, but would be darting across the face nonserially. We might just as well have an aperture for the numerals, for the circular face no longer serves any function. But in (b), even though the notation is syntactically differentiated, the serial clockwise arrangement of the numerals is a component of the symbol scheme—an analog component (the serial spatial ordering maps or models a serial temporal ordering). This component is missing from (c) and the difference between (b) and (c) is not appreciated in Goodman's explanation.[14] He might object that these cases are semantically undifferentiated, but the argument may work just as well for his dime counter. A circular dial with discrete numerals ordered serially, even if recording numbers of discrete objects like dimes, may still possess an analog component: the spatial seriation is an ordering which, even if discontinuous and differentiated, is analog to (maps) an increasing number of dimes. Even if the serial spacings are not proportionate to the increasing number of dimes, we know at least that a position further along the dial indicates more dimes (a tenuous but still significantly analog dimension of resemblance). Thus, Goodman's distinction fails to account for a significant representational difference—the fact that serial representations may still support a mapping relation, even if they are discrete and differentiated.

---

[14] It is important to notice, of course, that in *both* (b) and (c) there is still an analog correspondence between the serial ordering of integers and the sequence of time. My contrast here, therefore, restricts itself to the spatial medium.

Of course, there are counters whose elements are ultimately discrete, but whose unit of measure is considerably more indiscriminating. The pointer of a gas tank gauge cannot count individual hydrocarbon molecules. And so, there are innumerable systems which are *functionally* dense and undifferentiated. Lewis's example of a selector switch with each setting at 1-ohm could be taken as such an example.

Haugeland defends Goodman against Lewis. If Lewis's ohmmeter has a

> thousand positions, and the meter and resistors are good only to one percent, then (whether or not its analog) the device surely isn't digital; but it satisfies neither Goodman's condition nor ours. On the other hand, if there are only two hundred positions, and the meter and resistors are good to one part in a thousand, then it satisfies both Goodman's conditions and ours. But I think it is clearly digital—just as digital as a stack of silver dollars, even when the croupier "counts" them by height (*op. cit.*, p. 218).

We need only consider the two-hundred position switch Haugeland suggests. Both Lewis and I would hold out for significantly analog aspects of this representation against Goodman and Haugeland. Suppose we had a slide rule (a standard example of an analog device) whose slide could only be located above a scale in such a way that their markings would line up perfectly vertically (there would be "click-stops" for such alignments). If such a slide rule were manufactured as a "starter" for a seven-year-old, the markings might be quite far apart—an eighth of an inch or more—with no fine scale. An adult could interpolate between markings, but the instrument in the child's hands would probably be classified as digital by Goodman and Haugeland, despite the fact that the two linear serial extensions may still be mapped into one another to produce a desired result.

This mapping relation, I hold, qualifies the device as analog. It certainly still functions as a slide rule, and slide rules are prototypical analog devices.

Earlier, we found that discreteness is not a sufficient condition for the digital; now, we find that not even differentiation may be sufficient, for, on the model interpretation, differentiated representations may also be analog—as long as they remain *serial.* In a differentiated numerical series, every character can still be located between two others (ignoring the extremes); the gaps are neglected so that the series is treated as if it were dense, that is, the gaps either fall below some acceptable threshold of discrimination or are ignored. If a series exists whose gaps are relevant, for example, a television channel selector with click stops in UHF such that odd-numbered stops are missing even though the odd-numbered stations exist (that is, the latter forms a semantic compliance class, in Goodman's terminology), then the representation fails to some extent in its analog purpose, to be sure, but this does not ipso facto transform it into a digital one.

The discrete differentiated series we have considered as possibly analog, however, have been embedded in continuous series. The numbers ordered serially around a clock face are embedded in a circular spatial continuum (although the continuum is not part of the notation itself); even in the display of numbers in the window of a digital wristwatch, the serial ordering of numbers is taken to be an analog of the serial flow of time. The two-hundred position switch may also have its settings embedded in such a spatial continuum (assuming the switch is a dial or similar device). Even if the relative distances between the numerals on the circular clock face are unequal and therefore not exploited in the notation, the spatial continuum still serves to order the numbers serially.

We must consider the situation in which the syntactic medium is completely devoid of such embedding, where both syntactic and semantic media are discrete without either being embedded in continuous media. Many examples of this latter possibility can be formulated: one-to-one correspondences may be formed between the lifetime batting averages of all retired baseball players with equal numbers of the individual weights of all space satellites launched by the United States and Russia. Such a mapping could still be used to locate a desired result: if we find a very heavy American space station corresponding to Stan Musial's batting average, for example, we would know that, to discover the still heavier Russian payloads, we would have to look higher, toward Ted Williams. Because purely numerical series can thus function as maps for other such series, numerical *tables* may also constitute analog maps.

Despite the widespread contrary belief on this matter, therefore, even the use of numerical integers or purely mathematical notation may not be sufficient to determine that a representation is purely digital. The individual integers, like the sticks and balls in a model of a molecule, are obviously not analog to the individual elements to which they correspond, but the very fact that they may constitute a map indicates that some serial aspect of the original has been reproduced in the analog map.

Mappings between series are also legitimate examples of the analog in Shepard's account, mappings which establish what he calls *second-order isomorphism*. If individual mental representations (the neurophysiological events themselves) are isomorphic to their referents at the *first order*, Shepard explains, they would bear some structural resemblance to those referents.[25] My representation of a square must thus possess some structural similarity to the referent square; for example, the mental event actually contains four elements corresponding to the referent square's four corners.

A second-order isomorphism would not require such structural resemblance between individual mental events and their referents; rather, only a *relational* isomorphism would remain. The representation of the square would be judged to be intermediate between the representations of a triangle and a pentagon, for example, just as a referent square is intermediate between a referent triangle and pentagon; but the representation of the square need not resemble the referent square. Or, when I see a square, my mental event has something in common with past experiences of squares (*ibid.*, p. 89).[26] Shepard's analysis of the rotation of mental images supports "the idea that the representational process actually passes through a series of internal states corresponding one-to-one to an ordered series of intermediate orientations of the external object."[27]

Because Shepard's examples all involve spatiotemporal referents and because there is no reason to interpret the "mental states" that map these referents as linguistic or propositional in nature, it is difficult to know whether he would accept as analog a second-order isomorphism between nonspatiotemporal series. That purely numerical or descriptive series, as long as a mapping relation exists between them, qualify as analog, is the possibility I shall consider below.

---

[25] "Form, Formation, and Transformation of Internal Representations," p. 88.

[26] While Shepard claims that the perception of a square may involve first-order isomorphism, he denies that this isomorphism can be used to account for the capacity of the representation to refer; reference requires the kinds of patterns present only in second-order isomorphism (*ibid.*, p. 92).

[27] "Perceptual and Analogical Bases of Cognition," p. 53.

Pylyshyn finds value in Shepard's basic distinction, but uses the fact that even descriptions may form a series as a point *against* taking seriality as definitive of the analog. Pylyshyn gives the example of a description of the sequence of events which occurs when a batter hits a baseball (*op. cit.*, p. 203). It is not completely clear, however, whether he takes this descriptive "series" as a *logical sequence* of events (interpreted causally, for example), or merely as a *series of descriptions* matched one-to-one with the series of perceptions of the event. If it is a series of descriptions and not a logical sequence that Pylyshyn means, Shepard's criterion of second-order isomorphism is still satisfied.

Elliott Sober explores a similar interpretation, but not for similar reasons. He denies that the only possible interpretation of Shepard's image rotation account is one that relies on pictorial (analog) representation. A series of sentences can correspond one-to-one with a series of successive positions of the rotating figure, he suggests, so that each sentence differs from its neighbors only by the numerical value it assigns to some coordinate that specifies the angle for that position (*op. cit.*, p. 143). Thus, Sober observes, we do not have to attribute the reaction times that result to any actual mental rotation of images, but to the time it takes to run through a series of numbers.

Sober considers the possibility that analog representation exists even in a *series* of representations without continuously many members, but he rejects this idea on the grounds that even the series of purely linguistic descriptions that he offers as an alternative to an imagist interpretation would then have to be classified as analog (*op. cit.*, p. 144). Like Goodman and Haugeland, he clings to the sense interpretation.

Sober (and others) may resist construing the analog so broadly because their primary concern is in denying the imagists' claim that there is some mental version of rotation (with its own mental "space," and so on). But in being thus distracted by this aspect of the issue, they may be missing what may prove to be a more provocative possibility: that *serial* representations (or, as I shall argue, even those which establish only *relational identity*) may possess analog characteristics.

5. Relational Identity. Mappings between series involve an identity of *relation* between individual elements of the two series, but not an identity of individual elements themselves. Since such identity does not entail seriality, we have still, I believe, been too restrictive in our conception of the analog; that is, I propose that, because mapping requires relational identity, but not seriality, it is only the former which is necessary for the analog.[39] It is relational identity that is the basis for the model interpretation of the analog.

This opens up the analog to a number of additional types of relation. I want to outline a (rough) taxonomy of some of these possibly analog relations. I begin with some of the examples we explored earlier.

(a) Quantitative Analogs. We may distinguish single-variable and multivariable types. The analog relation between phonograph record grooves and the electrical impulses into which a phonograph cartridge converts this pattern is an example of a purely quantitative single-variable type. The relation between batting averages and satellite payloads (discussed above) is another example. Each instance of the given medium relates to neighboring instances according to quantitative variations of a single variable; this quantitative pattern of relation is reproduced in the analog medium. All simple tuners or tuning mechanisms and most elementary measuring devices rely on such a simple analog relation (knobs, dials, gauges, clocks, and the like).

More complex tuners and maps may employ multivariable quantitative analogs. Two spatial coordinates can be used to represent two variables in an analog medium—the tone and intensity of a sound, for example. Thus, each point of this two-dimensional space corresponds to a pair of values (tone, intensity). Two- and three-dimensional spatial coordinates are often used to depict analog relations in other media; this is the utility of all graphs. Maps and complex physical and mathematical models also embody such analogies: each state (instance) of a model bridge involves specific values of many relevant variables relating to stresses, loads, and so on; these states (it is hoped) form one-to-one correspondences with states of a full-scale bridge. The model is a multidimensional (or multivariable) analog of the original.

Quantitative single-variable cases can only establish analog relations between multiple *values* of the single variable in each medium.

---

[39] If one were to characterize the two relations in the example I used earlier as "*A* is a heavier payload than *B*" and "*C* is a higher batting average than *D*," then it is obvious that the relations per se are *not* identical. The "identity" to which I refer, therefore, exists between more formal aspects of the two relations, namely, their purely quantitative aspects. That is, to use K. Holyoak and Paul Thagard's terminology, even when the relations are largely dissimilar, a *structural consistency* exists— "Analogical Mapping by Constraint Satisfaction," *Cognitive Science*, XIII (1989): 295-355. This is also the sense of identity in Shepard's isomorphism.

Once additional variables are introduced, however, relations between *variables* of a single medium yield a new basis for isomorphism or relational identity. Analog relations may thus be established between (i) a "structural" model and its original and between (ii) two rules.[29]

(i) The example of a model bridge provided above suggested how the relations between individual states of the model might be identical to the relations between individual states of the original (where a single "state" would comprise a set of specific values of multiple variables); but this model can be read in another way. That is, we can take the relations between the *variables* of the model bridge to be identical to the relations between the variables of the original. It is not a series of states of the bridge which is analogous to a corresponding series in the original, but the *structure* of the bridge which is analogous to the structure of the original. The term 'structure' suggests some theoretical understanding of the bridge—of the sort one would expect in mathematical knowledge; but we may speak of a structure even in the absence of a knowledge of the theoretical relation among the variables. The individual essential variables of a given phenomenon may be completely articulated, but their theoretical relation (the digital rule that abstractly relates these variables) may be unknown; the variables are therefore left *in situ*, that is, they are viewed as concretely embodied in a typical instance. There is a distinct sense of "model" which is appropriate for this concretely exhibited relation among articulated variables; such a "model" would amount to an intermediate state of representation, possessing both analog and digital characteristics—quite like the example of the carbohydrate model, where individual structural parts are digital symbols, but where the relations remain spatiotemporal.

(ii) A theoretical relation among variables (as distinct from the concrete "display" of these variables in a model) constitutes what I have elsewhere called a cognitive *rule* (*ibid.*, ch. 11). It is the notion of rules, in fact, which I believe to be the appropriate basis for understanding the nature of *digital* representation (space does not permit my development of that idea here). Coulomb's law—that the force with which two electrostatically charged spheres repel or attract one another is equal to kQ_1Q_2/r^2—is a rule accounting for the set of all instances of values satisfying the rule: (Q_{1n}, Q_{2n}, r_n). While Coulomb's law is a rule accounting for its extension, however, it may also be taken as an analog model for the rule: gravitational force = Gm_1m_2/r^2; for while the individual elements are not identical, their relation is.[30]

---

[29] I provide an account of the differences between models and rules and their role in a theory of representation in my *Of Two Minds: The Nature of Inquiry* (Albany: SUNY, 1997), chs. 10-12.

(b) Qualitative Analogs. Relational identity requires genuine *relations.* In quantitative analogs, relations may exist between *values* of a single variable (for example, numerical ratios) or between *variables* themselves (for example, in Coulomb's law); these relations may be duplicated in an analog without duplicating the individual relata. In qualitative cases, however, while relations between variables exist (examples to be provided below), it is more difficult to conceive of a relation between *values* of a variable.

Consider the analogy: a cypress is to a cactus as a water moccasin is to a sidewinder. While this may look like an analogy founded on a relation between two values (tropical and arid) of the respective variables (plants and snakes), a quantitative basis may be surreptitiously intruding, for tropical and arid climates can be treated as two values of a quantitative scale (of temperature), with intermediate degrees as well. When there is no quantitative basis at all for the comparison of two values, we sometimes say it is a case of "apples and oranges." Suppose I say: Grace Kelly is to Charles Bronson as a swan is to a bear; or Mozart is to Beethoven as Kant is to Hegel (you will need to stipulate the accuracy of such analogies for the purposes of discussion). Here, the differences between the values which yield the different instances seem to be simply differences of *kind,* not of degree or contrariety. But because the "relation" of beauty to strength which is the basis for the first example is arbitrary, so that we might well ask whether this constitutes any genuine relation at all, there is a serious question whether this should qualify as a true analog.

It is a much easier task, of course, to find qualitative analogs between two theoretical relations (rules) and between a structural model and its original, rather than between two sets of values, for it is here that much of ordinary analogy exists. The president of the country is the captain of the ship. This analog claim is based on the relational identity between the two forms of leadership. It is the qualitative version of the rule/rule analogy between Coulomb's law and the Newtonian law of gravitation.

Mapping entails relational identity. The examples of mapping I have provided have all been of the quantitative type, where the use of a map relies on our perception of proximity, intermediacy, and so on—all quantitative judgments. But there are qualitative maps as well. Suppose I construct an analogy between the conflict observed between two political parties and the conflict between two football teams. The chairperson of the party may be taken as the quarterback, the offensive line as the chairperson's coterie of protective aides, and so on. Now, suppose someone challenges me to find an analog for disinterested parties, such as the judges and magistrates representing the legal system. I would be able to use the analogy as a map and locate a corresponding element within the football model—the referees, judges, and umpires, for example.[11]

---

[30] In fact, the discovery of Coulomb's law was taken by some as a vindication of Newtonianism because of this resemblance.

If even nonserial and nonquantitative conceptual representations support some degree of relational identity with their referents, then it will be possible to conceive of the analog without either ascribing it to purely physical constraints (as Pylyshyn does) or having to settle the question of the nature of mental images: a scaled-down physical model, a mental image, and an abstract analogy may all be analog.

II. THE ANALOG BEYOND MENTAL IMAGERY

Let us now return briefly to the debate on mental imagery. I have already suggested that it was wrong to conceive of the relation between analog and digital forms of representation in terms of the relation between pictorial images and linguistic or propositional descriptions (or to conceive of the latter relation in terms of the former). These two distinctions may not be co-extensive because, as I have argued, there may be analog mental states that do not rely on the spatiotemporal media essential to imagery. Consider again Kosslyn's characterization of imagery. Does he take spatiality to be a necessary component?

Depiction in imagery, Kosslyn observes, relies on "quasi-pictures," not literal ones. What is the difference? He contrasts the "physical space" in which literal pictures are represented with the "functional space" in which images are formed:

> The space in which the points appear need not be physical, such as this page, but can be like an array in a computer, which specifies spatial relations purely functionally. That is, the physical locations in the computer of each point in the array are not themselves arranged into an array; it is only by virtue of how this information is "read" and processed that it comes to function as if it were arranged into an array (with some

---

[11] Another important framework in which qualitative mapping may occur is provided by trees of logical classification. An apple is to a fruit as a carrot is to a vegetable. Here, the relation is one of logical inclusion. But this entails quantitative features as well: "apple" is (quantitatively) "closer" to the class "fruit" than to the class "plant."

> points being close, some far, some falling along a diagonal, and so on) (*op. cit.*, p. 5).

Elsewhere, Kosslyn explained that "*distances* in the medium can be defined without reference to actual physical distance but merely in terms of the number of locations intervening between any two locations."[22] The question is: Is such a *functional* space still spatial? Could we treat the set of points as *serial*, but no more spatial than the series of whole numbers? If we could, then Kosslyn's conception of the "pictorial" (analog) may not entail spatiality; rather, a spatial array would be an "interpretation" or instantiation in space of the serial array. This would be, finally, a step away from the sense interpretation of the analog. But why then call it functionally *spatial* at all? An innocent answer would be that Kosslyn is, after all, concerned with a type of cognition—imagery—that seems to be inherently spatiotemporal. Would he allow for the possibility of such serial arrays in completely nonspatial cognition? As far as I know, he does not address this question specifically,[23] but he does give the impression that all representation falls on one side or the other of the depictive/propositional division; and my point is that nonimagistic analog cognition is *both* nondepictive *and* nonpropositional (in his senses of those terms). (Of course, I would also hold that such analog cognition is not confined to *serial* displays, but is found even where we establish only relational identity.)

III. ABSTRACT ANALOGS

It is Pylyshyn's view that analog models are reducible either (i) to features of the physical functional architecture or (ii) to purely computational operations or descriptive rules or (iii) to a combination of these. That is, there is no place for models as analog mental states. Yet it is precisely as *models* that analog representations first came to be distinguished as such (in the notion of the "analog computer"). It would be particularly important in any effort to defend the model interpretation of the analog, therefore, to have a theory of purely *abstract* models, for it would be difficult to explain these away by an appeal to the physical mechanisms on which perception (and possibly imagination) rests. Further, the very existence of analog yet abstract mental models would throw into question the plausibility of the sense interpretation of the analog.

---

[22] "The Medium and the Message in Mental Imagery," in Block, ed., *Imagery* (Cambridge: MIT, 1981), pp. 207-44, here p. 215.

[23] In reviewing Pylyshyn's and Sober's analysis of the series of descriptions that might correspond to a series of perceptual states, I pointed out that they had both missed an opportunity for constructing a broader and more fruitful conception of the analog—where even a serial ordering of pure descriptions would possess significantly analog characteristics. In discussing Kosslyn's view of the arrays necessary for imagery, Michael Tye points out (reflecting both Kosslyn's and his own view) that, should the elements of an array be descriptive rather than pictorial, no functional mental image would be involved—*The Imagery Debate* (Cambridge: MIT, 1991), p. 40. Like Kosslyn, Tye does not consider the broader possibility that what is essential to analog forms of representation is not the spatiality of the array but the array itself.

Philip Johnson-Laird[14] is one of a few cognitive theorists to take the debate on the nature of the analog out of its traditional context in the discussions concerning perception and mental imagery and defend the view not only that analog models exist in purely abstract cognition, but also that they are genuinely nonpropositional in nature. He writes:

> Whenever I have talked about mental models, audiences have readily grasped that a layout of concrete objects can be represented by an internal spatial array, that a syllogism can be represented by a model of individuals and identities between them, and that a physical process can be represented by a three-dimensional dynamic model. Many people, however, have been puzzled by the representation of abstract discourse; they cannot understand how terms denoting abstract entities, properties or relations can be similarly encoded, and therefore they argue that these terms can have only "verbal" or propositional representations (*ibid.*, p. 415).

Johnson-Laird holds that nonpropositional mental models are used in purely abstract reasoning, but it is not clear exactly how he distinguishes abstract mental models from those which rely on imagery. Let us see what he says more specifically about this difference. He explains that

> The psychological core of understanding consists in your having a "working model" of the phenomenon in your mind. If you understand inflation, a mathematical proof, the way a computer works, DNA or a divorce, then you have a mental representation that serves as a model of an entity in much the same way as, say, a clock functions as a model of the earth's rotation (*ibid.*, p. 2).

It is fairly apparent that Johnson-Laird requires a relational identity between a model and what it models (*ibid.*, esp. p. 403). He insists, for example, that a model relies on "symbolic notation"; that is, that while relations between elements in the model may resemble relations between elements in the original, an individual element of the model does not resemble an individual element of the original, the former only "symbolizing" the latter (*ibid.*, p. 404).

---

[14] *Mental Models* (Cambridge: Harvard, 1983).

Johnson-Laird acknowledges that perception is a primary source of mental models and that higher uses of models are a "natural extension of their perceptual function," but he distinguishes models from images. Images, he says, "correspond to *views* of models," and he relates images to models in much the same way David Marr relates two-and-a-half-dimensional sketches to three-dimensional representations (*ibid.*, pp. 157, 406): a model, like Marr's three-dimensional representation, is not a perceiver-centered view, but a "God's-eye view." Models are like Kantian *schemata* and he affirms (with Kant) that "a schema is not an image, but a model that underlies the ability to make an image" (*ibid.*, p. 190). Nevertheless, models are not propositional, that is, composed of "strings of symbols." But how does he treat purely abstract cases?

An abstract concept without a mental model, he observes, is a concept without foundation, a purely propositional representation; but "as long as individuals have clear intuitions about how the world would have to be for an assertion to be true, they must have transcended a purely propositional representation" (*ibid.*, p. 415). In some way, mental models stand in for the world and function as the referent for propositional representations. Because discourse may also be about fictitious worlds, building models "frees us from the fetters of perceptual reality" (*ibid.*, p. 407).

Yet much of this is consistent with a view of models as analog images. Johnson-Laird recognizes that, insofar as a mental model derives from perception, "it will be a single entity corresponding to a single perhaps unique state of affairs" (*ibid.*, p. 407), whereas discourse is considerably more indeterminate. His solution is to propose that "a mental model can stand for an infinite number of possible states of affairs, because the model can be revised recursively"—a revisability that itself depends on the mapping the model provides. That is, a model represents a typical state of affairs and is altered only when conflicting instances arise. Such is the nature, Johnson-Laird argues, even of the concrete "psychological logic" employed in actual syllogistic reasoning.[*] We learn new concepts and refine old ones through such procedures, so that we may trace a trajectory for a given concept over a series of its models—each of which may prove more adequate for the extension of instances it encompasses.

So far, Johnson-Laird's conception of a purely abstract model is at least consistent with the account of the analog I have proposed. It is especially significant that he takes a model's editability (made possible by the mapping or relational identity established between itself and its referent), rather than any imagistic content, to be its essential characteristic.

---

[*] W. Bechtel and A. Abrahamsen have made the same point, but without reference to Johnson-Laird's analysis; see "Beyond the Exclusively Propositional Era," *Synthese*, LXXXII (1990): 223-53, esp. pp. 247-48.

Yet my attempt to disengage the question of the nature of analog "representation" from the imagery debate would be in jeopardy if it were true that *all* analog cognition—including that found in mathematical and purely abstract reasoning, for example—in fact relied on some sort of spatial imagery. A favorite example of spatial-image-based reasoning is that in which some quantitative difference among distinct individuals (height, intelligence, and the like) is converted into differences in location on a spatial line. It is easier to understand the relations among the propositions 'Sally is smarter than Sam', 'Sam is dumber than Sheila', and 'Sheila is dumber than Sally', for example, Kosslyn suggests, by employing a spatial representation of these relations.[26] Yet while the utility of spatial media in abstract reasoning should not be underestimated, there is another perhaps more basic process of which this usage may prove to be a special case: I mean the employment of concrete *instantiation* in problem-solving strategies. There is evidence that reasoning does not occur (necessarily at least) through the use of completely abstract formulas and descriptions[27]; and also that categorization may rely less on the application of abstract concepts definable through necessary and sufficient conditions and more on the use of concrete prototypes plus some editing procedure.[28] In both circumstances, we apparently find it easier to concretize or instantiate some abstract process by working with typical or "dummy" values of variables than to eschew instantiation altogether and calculate results solely by means of algorithms that determine abstract relations among variables. It is also interesting that this principle may apply even to the lower-level processes involved in imagery, so that as Michael Tye argues, for example, our image making requires instantiation by means of a two-and-a-half-dimensional sketch rather than remaining at the level of three-dimensional models (*op. cit.*, p. 89).

Does instantiation require spatialization? I have argued elsewhere[29] that it does not—that we are quite capable of instantiating abstract concepts, such as "number" or "concept," for example, without invoking spatial media; such instantiation still amounts to a digital-to-analog conversion. Johnson-Laird seems to have a similar view. He recognizes, for example, that "the analogical structure of mental models can vary considerably. Models of quantified assertions may introduce only a minimal degree of analogical structure, such as the use of separate elements to stand for individuals" (*op. cit.*, p. 156); that is, instantiation without any reliance on spatiality. Even though he does not pause to make a point of this crucial distinction, he has by his broadened discussion of mental modeling helped to transpose the discussion of the analog from its locus in the debate on mental imagery to a larger context where it serves to anchor the problem of nonpropositional mental states in general. Of course, we could still speak of imagery in such abstract cases if we took "imagery" to refer to instantiation per se without any spatial component; but we would then, I think, be bending the word toward a metaphorical use.

---

[26] *Image and Mind*, p. 324; see also Tye, p. 60.

[27] As Johnson-Laird argues in *Mental Models*, for example.

[28] As E. Rosch's work suggests; see "Principles of Categorization," in Rosch and B. B. Lloyd, eds., *Cognition and Categorization* (Hillsdale, NJ: Erlbaum, 1978), pp. 27-48.

[29] See my "Unarticulated Meaning," *Erkenntnis*, XL (1994): 43-70, and *Of Two Minds*, ch. 12.

It is tempting to view purely abstract analog mental states (and equivalent mental models) as, simply, *meanings*[8]—at least in a psychological sense. This is implicit in some of Johnson-Laird's discussion of the role of models in establishing contact with the world, and in his account of concept learning: "When you teach someone the meaning of a word, you describe examples of the range of its application" (*op. cit.*, p. 189). Our grasp of a given meaning may be inarticulate and vague, in contrast to the rule-like propositions that come to articulate it; yet even unarticulated meanings may be concretely grasped and manipulated. It is when models and meanings are overloaded, when their richness cannot be sustained without dissipating, that a propositional rule enters that can assume the burden of that complexity. On the other hand, it is when such rules prove incomplete or inadequate—

[8] The fact that meanings seem to be grasped before they come to be expressed has sometimes led to their association with analog images. Sober relates Gottlob Frege's denial that mental images can form the content of the meaning of terms on the grounds that images (unlike given meanings) vary greatly from person to person, but Sober judges that images need not be taken to be so variable, a judgment based, however, on his view that images are really "theoretical procedures" (*op. cit.*, pp. 102-03). Fred Attneave holds that images may be reconstructed in an analog mode from a nonanalog description in which they are stored and suggests that the meaning of these nonanalog descriptions may simply be the imagery they can generate—"How Do You Know?" *American Psychologist* (1974): 493-99, here p. 498. Shepard discusses the internal events corresponding to meanings in terms of representations that are neither images nor verbal in nature—"Form, Formation, and Transformation of Internal Representations," pp. 110-11.

when, that is, we encounter the limits of representation—that we return to meaning. This is not, however, an idea that can be pursued here.[41]

IV. CONCLUSION

As an alternative to the sense interpretation of the analog, which would require analog representations to be continuous or at least dense, I have in this discussion promoted the model interpretation: that the function of analog representation is to map or model what it represents. My argument disqualifies *none* of the examples of the analog on which the former interpretation relies; rather, it expands the notion of the analog to include many more types.

I have also tried to exploit what might be taken as the analogy between the analog and analogy itself. Clearly, there is *something* common to the traditional notion of the analog as a modeling or mapping representation and analogy, namely, relational identity. While in certain circumstances it might be practical to use the sense interpretation of the analog (in discussions regarding the nature of imagery, for example), it is the model interpretation, I maintain, that is a more principled conception of the analog and (therefore) more significant for a general theory of representation.

It might be objected that, if relational identity is chosen as the basis for the analog, the analog is thereby trivialized, for it would now be so broad so as to include everything. It is true that analogy is an exceptionally broad notion: by choosing just the right aspects on which to base an analogy (that is, on which to establish a relational identity), it may well be possible to strike analogies between any two things. Thus, "a chimpanzee is like a tennis ball": this is a challenging assertion, sounding rather like a puzzle or semipoetic exaggeration. Are chimpanzees bouncy? Or are both fuzzy? A less clever and more straightforward response could simply be: both have gravitational mass. Analogies such as this may not be very useful, of course, but they qualify as instances of the analog. The analog does not become a vacuous concept because it is so comprehensive. Everyone understands that analogy is an option—that we can just as easily concentrate on the differences between nonidentical items and ignore any of their analogous aspects. When such a shift occurs, however, the manner of representation itself changes. This claim is only half-supported, since I am not here able to provide a corresponding analysis of digital representation, which would show more clearly what it would mean *not* to represent something as an analog.[a]

---

[41] I treat meanings (in this psychological sense) as nonrepresentational mental states in "Unarticulated Meaning."

Those who prefer to identify the distinction between analog and digital representation with some version of the traditional division between sense and thought would also have to explain away the apparent exceptions that my own proposal would more naturally account for. Apparently, analog aspects of thinking (including abstract analogy), they would have to claim, either are not analog at all (but are explicable in terms of digital representations) or are attributable to the intrusions of perceptual imagery. Correlatively, apparently digital aspects of perception either are not digital (but explicable as some form of analog intuition) or attributable to the intrusions of thought and language. I have suggested, on the contrary, that we examine the common supposition supporting this reductionism (prevailing since Kant) that sense without conception is "blind" while conception without sense is "empty." If the analog-digital distinction cuts across rather than is parallel with that between sense perception and thought, a new model is suggested—where thought possesses its own analog content (as I have suggested), and where sense is guided by its own digital rules. That is, both analog and digital mental states would be found throughout our cognitive life, from the simplest perception to the most complex forms of abstract thought.

JAMES BLACHOWICZ

Loyola University of Chicago

---

[a] The term 'representation' has broad and more narrow uses. Elsewhere (*Of Two Minds*, chs. 10-11), I argue (in agreement with Pylyshyn and Fodor) that "analog representation" (taking 'representation' more narrowly than I have done here) is indeed impossible—that representations in the strict sense must be rules for analog extensions (this establishing their digital nature) and that relational identity is used in lieu of (or en route to) full representation. That discussion completes the one offered here.

COMMENTS AND CRITICISM

ELGIN ON LEWIS'S PUTNAM'S PARADOX*

In "Unnatural Science," Catherine Z. Elgin[1] examines the dilemma that David Lewis sees posed by Hilary Putnam's model-theoretic argument against realism. One horn of the dilemma commits us to seeing truth as something all too easily come by, a virtue to be attributed to any theory meeting relatively minimal conditions of adequacy. The other horn commits us to "antinominalism," some version of the ancient doctrine that language must "carve nature at the joints": that there are natural kinds or classes that alone qualify as referents (extensions) for our predicates. Elgin offers a searching critique of Lewis's response (accepting the second horn) and an illuminating defense of its contrary: "we cannot construe (mere) truth as the end of scientific inquiry. Not...because truth is too hard to come by, but because it is too easy" (301).

I have no issue to take with Elgin's arguments, or indeed most of her main conclusions. But she grants Lewis too much. When Putnam's argument was construed as posing that crucial dilemma, it became an entrance into the very kind of metaphysics Putnam intended to undercut. First, I shall outline exactly how (as I see it) Elgin follows Putnam and Lewis in their dialectic. I mean to be accurate but not slavish: the outline should bring to light some possibilities of opting out at various points along the way.

I

Elgin's statement of Putnam's argument brings out very clearly its initial incredible degree of generality. Assume the world consists of infinitely many things and consider any theory T that says as much.[2] Then there is an interpretation of the language (call it L) in which T is stated which assigns, to all its terms, extensions in the world and which satisfies T. Neither Lewis nor Elgin nor I demur from the proof by which Putnam establishes this.

Elgin completes the argument: "Under the resulting truth definition, truth (under that interpretation) is truth. (T) then is true" (289). A moment's pause: What explication of truth is indicated here? Let us give it a name:

---

* I want to thank Catherine Z. Elgin for very helpful discussions and correspondence; also Igor Douven, Mary Kate McGowan, and Gideon Rosen for helpful comments.

[1] This JOURNAL, XCII, 6 (June 1995): 289-302.

[2] However we construe this 'says as much', it implies that T is consistent and has an infinite model.

0022-362X/97/9402/85–93

(EQ) *T* is true if and only if *T* is true under an interpretation that assigns parts of the world as extensions to the terms of *L*.

The first response will be to reject this as too generous.

II

Putnam himself had restricted the scope of his conclusion to ideal theories, as Elkin acknowledges: *T* "answer[s] to all observational evidence, past, present, and future. It is theoretically adequate: consistent, simple, informative, and so on (add any other theoretical virtues you like)" (289).[3] These qualifications play absolutely no role in the proof to which, as I said, we offer no demurral. What role do they play, then?

In (EQ) we see all possible interpretations admitted on a par, even the most gruesome (even those which interpret 'green' as 'grue') in Elgin's elegant turn of phrase. But if *T* can be a real theory stated in part of our language, many such interpretations are *unintended*. The restrictions on *T* may be regarded as means to cull interpretations. This insight of Lewis's shapes much of the ensuing discussion. That all observation reports framed in *L* and never taken back or formally contradicted, in the history of the speakers of *L*, must come out true may be taken as a constraint on interpretations. If *T* is assumed to contain all those statements, that will ensure that, if it is satisfied by any interpretation, it is satisfied by one that meets this constraint. Therefore, the restriction on *T* may be seen equally as a constraint on the *admissible* interpretations (candidates for the status of "intended" interpretation).

Putnam himself appears to hold that any constraint we may use to cull interpretations can take this form: the form of specifications concerning what an ideal theory is to contain. That is why he says that any constraints we can impose are all "just more theory."

> If Putnam is right, there is nothing *we can do* to evade this result. For whatever we do is just more theory. I think he is right. So does Lewis (290).

The reflection that the original proof was insensitive to the content of *T* is, of course, grist to Putnam's mill. We may replace the explication of truth so as to accommodate the intended/unintended distinction, by revising (EQ) to:

---

[3] I have omitted Elgin's addition of 'empirically adequate'. As I understand it, a theory is empirically adequate if it is true in part, to a certain extent. The virtues that make for an ideal theory must, in this context, all be ones that we can in principle (if only in the ideal long run) tell before facing questions of truth. Otherwise, why not include truth overall in what makes for an ideal theory? I have also, like Elgin, omitted 'complete', which occurs in Putnam's text but seems neither to play much of a role in the argument nor be the sort of desideratum we could humanly demand of our theories, even in principle.

> (EQ*) *T* is true if and only if *T* is true under an *admissible* interpretation that assigns parts of the world as extensions to the terms of *L*.

Suppose all the culling of interpretations can proceed, in principle, by insisting that (on any intended interpretation) certain classes of statements in *L* are to come out true. Require ideal theories to contain all those favored statements. Then the original proof still works, and we conclude:

> If an ideal theory is true on (EQ) then it is also true on (EQ*).

This defeats any possible critique based on Putnam's mistakenly granting equal rights for the linguistically disadvantaged, in such matters as these.

III

Enter Lewis's dilemma. The first horn: every ideal theory is true (on some admissible interpretation—hence, by EQ* or similarly strong principle, true). Second horn: to be admissible, an interpretation must have virtues beyond that of meeting our intentions. The latter is possible, on the face of it. As Elgin phrases Lewis's second insight, "even if *we* cannot disqualify motley collections from serving as extensions of our terms and truth makers of our theories,...the world may disqualify them for us" (290). There may be privileged or "natural" classes that alone may be assigned as extensions; any interpretation that fails to do so is inadmissible. Elgin's main agreements and disagreements with Lewis are concisely stated at exactly this point. Using 'Putnam's result' to denote the conclusion that every ideal theory is true (see what I called the first horn of the dilemma), she writes:

> Lewis contends that unless the world privileges certain schemes of organization, and precludes reference to sufficiently gruesome properties, Putnam's result is inevitable. I agree. Lewis concludes that the world confers such privilege. I conclude that Putnam's result holds (290-91).

This is also the very point where, I shall argue, Elgin has gone too far to meet Lewis's terms for the dialectic.

IV

The insertion of 'admissible' into the truth conditions for a theory certainly did something. It provides the form of a solution to the problem posed by the initial argument. Form is not enough. The putative result that all (ideal) theories are true can be avoided by a solution of this form, but only if it has something to be said for it in its own right. Thus, Lewis faces two tasks. The first is to give real content to the distinction between natural and arbitrary classes (how the world can "privilege" certain classes or properties). 'Admissible' cannot stand for something like 'whatever it takes to avoid the unwanted result'; it needs real content.

The second task is to provide support for the claim that natural classes are the sole candidates for reference. I oversimplify: natural classes are the prime candidates, and anything defined from them is also a candidate, to a degree that varies directly with the complexity of definition (however that is to be spelled out). The simplification does not affect the challenge very much; both the distinction between natural classes and the character or nature of reference need to be spelled out satisfactorily. That means: in such a way that there will (or at the very least, can) be evident support for the claim that reference in our language, by means of our terms, is restricted in such fashion.

Elgin's main arguments are to the effect that this second task cannot be carried out successfully, regardless of how the first is performed. To me, these arguments seem very telling, but I shall leave their examination to Lewis; I am anyway not a disinterested party in this matter.[4] Nor do I mean to examine the possibilities of success for the first task. Rather, I mean to examine the grounds on which these two tasks could become imperative in the first place.

V

In the initial argument, when Elgin completes Putnam's argument by drawing a devastating conclusion from his proof (the conclusion Lewis called Putnam's "bomb"), I construed her as relying on an explication of truth conditions for theories:

> (EQ) T is true if and only if T is true under an interpretation that assigns parts of the world as extensions to the terms of L.

Apart from cardinality constraints due to the world's size,[5] this equates truth with consistency—which is absurd. Not only absurd: that equation allocates truth to distinct theories that stand in formal contradiction to each other, thus engendering inconsistency in the strict formal sense. Thus, if we like, we can take the initial argument simply as reductio ad absurdum of this simplistic attempt to relate the notions of truth, on the one hand, and truth in a model or under an interpretation, on the other.

---

[4] See my *Laws and Symmetry* (New York: Oxford, 1989), ch. 3, sect. 5.

[5] So to speak. What matters is the finite/infinite distinction; and that the context of discussion is classical rather than intuitionistic.

The revised condition (EQ*), which results from inserting the placeholder 'admissible', is not subject to that quick reductio. But that is, initially, only because of our ignorance about what is and is not admissible. What must supplant the placeholder 'admissible' to put a genuine halt to Putnam's argument?

If the language has two genuinely differing interpretations, there must exist two distinct complete theories (possibly framable only in extensions of language *L*) which are satisfied by those two interpretations. Given (EQ*), they are both true. But being distinct and complete, they stand in formal contradiction to each other. Although their conjunction may not be expressible in the language in which they are stated (namely, if they are not finitely axiomatizable), it remains that the law of noncontradiction is definitively violated. That is an absurd consequence. Thus, (EQ*) is reduced to absurdity in exactly the same way, unless there is only one single unique admissible interpretation of the language.

I am by no means protesting vagueness of reference or the idea that candidacy for reference is a matter of degree. On the contrary! I would assume these to be features of any realistically conceived language. But if they are present, or if even the least bit of leeway of any other sort is allowed, then the language will have more than one admissible interpretation. If that is so, the acceptance of (EQ*) throws us immediately into inconsistency.[6]

Escaping the import of Putnam's argument by the insertion of 'admissible' looked easy. The insight that we can circumvent Putnam's own very limiting strictures, on how intended interpretations are to be demarcated, may have increased its appearance of feasibility. But the ease is deceptive. The reductio is avoided only if correct interpretation is unique—that which there may be very little, in the entire history of philosophy, more difficult to defend. And this on pain not of skepticism or metaphysical discomfort but outright logical inconsistency.

It may not be amiss, at this point, to remember Donald Davidson's longstanding cautions about simplistic links between truth and relativized truth. We cannot equate ' 'Snow is white' is true' with 'there is a model of type *X* in which 'snow is white' is true', unless we can clearly show that the latter is the case if and only if snow is white.

---

[6] To postulate or require that real language cut nature at the joints gives us a solution that (we granted) does indeed have the form of possibly tenable solution. But look at how much weight the postulate must bear. Suppose that the only constraints we can put on interpretation is that certain sentences about cats, accepted during the history of the human race, must come out true. Suppose in addition that the only further constraint upon interpretation is that 'cat' must have as extension a natural kind or class. Between the two of them, can they really select a single candidate for that extension?

VI

> When interpretation is our aim, a method of translation deals with the wrong topic, a relation between two languages, where what is wanted is an interpretation of one (in another, of course, but that goes without saying since any theory is in some language).[7]

Why should something like (EQ) be so appealing, when it leads so quickly and almost unavoidably to inconsistency? Well, if *T* is stated in a foreign language, an artificially created language, or a part of our own language that we place under study (during which we describe but do not use it), something like (EQ) is clearly correct. That is, in such a case we are allowed to regard *T* as true if we can find an interpretation which answers to all our constraints on interpretative engagement, and under which *T* is satisfied. But note what is allowed, even then: to take *T* to be true *as construed*.

To go beyond what I have just delimited (either by not restricting the theory under discussion to one stated in described rather than used language, or by allocating truth simpliciter under such conditions) is to bring us at once into those semantic paradoxes to which Alfred Tarski himself alerted us.

The point is easily obscured, for it is only right that, for example, English should be studied by English speaking linguists, natural language by those with natural linguistic competence, brains by those who have brains, intelligence by the intelligent. However crudely I have managed to state it now, the point is not only easily obscured but in some of its aspects exceedingly subtle. Consider the question asked by a Dutch speaker:

> Is het Engels woord 'green' een woord voor groene dingen?

and its putative English translation:

> Is the English word 'green' a word for green things?

Despite the extreme care I took in translation, recalling even how Alonzo Church famously rapped Rudolf Carnap's knuckles, the two questions are not the same, at least not for native speakers of those two languages. The first question is an empirical question, which the Dutch speaker must settle by empirical research, field work, and can settle to the extent that scientific research can settle any empirical question. But the second question, posed or heard by an English speaker, calls for no empirical research. For the English speaker to contemplate a negative answer thereto is to lose coherence. This despite the obvious consideration that the English language might and indeed could have evolved so as to make 'green' a word for red things—and might yet evolve into that condition. But though easily obscured, and in some of its aspects exceedingly subtle, the point is not to be missed.

---

[7] Donald Davidson, *Truth and Interpretation* (New York: Oxford, 1984), p. 129.

VII

Suppose I do want to explicate the formal conditions of truth for an actual scientific theory—say, QED or psychoanalysis—stated in English. To see whether it is true under some admissible interpretation I must be able to describe the domain of those interpretations (that is, the part of English in which that theory is stated) and their range (that is, the subject matter of those theories). I shall do both in my own language. Indeed, I cannot otherwise. But now, on this assumption that the theory is stated in the language I am using, I shall court contradiction, unless I insist that the admissible interpretations are those which assign the set of photons as extension to the predicate 'is a photon', the set of green things to 'green', and so forth. Thus, we have not only an immediately accessible way to *disqualify all unintended interpretations*, but one which is imperative, which we must use and may say nothing to contravene, on pain of inconsistency in our own mouths.

Am I willfully and unphilosophically dismissing skeptical doubts when I say this? Do I fail to feel the force of those qualms about relying on a language whose reference we have not demonstrably "fixed" firmly in the world? (Do we understand this metaphor?) Just what doubt am I supposed to harbor? The doubt that some green things are not in the extension of my predicate 'is green'? Or the doubt that some things that are not green are accidentally included? Either doubt would make me say things contrary to:

> The extension of 'is green' is the set of green things.

and so equally contrary to:

> 'so and so is green' is true if and only if so and so is green.

thus placing me in formal contradiction with some of my most basic convictions, sure to be reflected in any views I propound elsewhere about truth and reference.

VIII

Having listened to Tarski and Davidson, to what shall we liken the "carve nature at the joints" solution to Putnam's paradox? Let us listen to a dialectic that could lead to that solution.

Imagine someone struck by the point noted above, that the "*T*-sentences," such as:

'Snow is white' is true if and only if snow is white.

which are to be among the consequences of any adequate theory of truth (for the relevant part of our language), are not logically true and do not express necessary propositions. The English language could have developed in such a way as to make 'white' a word for black things, and might so evolve yet. Having listened to Tarski, this person is convinced that the *T*-sentences must all be true; but how to explain that?

Worse: if 'white' could have had the set of black things as its extension, how can we be sure that it does not? Since snow is white, if 'white' were a word for black things, then 'Snow is white' would not be true. Is there anything we can do to secure guaranteed truth for all the *T*-sentences? That would require us to get all the extensions of all the predicates "fixed" in just the right way. There is nothing *we can do* to make sure of this, with so many classes out there as candidates for the extension of each predicate.

But wait: reference is a two-place predicate. If we cannot do it alone, the other relatum can help. Our skeptical doubt, which sees the possible extensions of 'white' ranging over all the classes of things there are, is alleviated if we postulate that nature itself restricts the candidates for its extension. Not removed, perhaps, unless we add the hope or postulate of uniqueness; but certainly alleviated.

What is wrong with the reasoning that invited this "solution"? Our imaginary person is in effect trying to play two roles at once. To raise a doubt about whether 'white' is a word for white things, or whether 'Snow is white' is true if and only if snow is white, *in these very words*, requires two things. To do that, on the one hand, she must use the words 'white' and 'snow' to talk about white things and snow; on the other, she must treat the quoted expressions as if they belonged to a foreign language. To do the latter is to insist that 'white' is part of 'Snow is white' only in the sense, famously, that 'can' is part of 'canary'. The first role is that of native English speaker. It accounts for the felt need to make the *T*-sentences come out true at all costs. The second role is really the same as that of the Dutch speaker who asks (in his own language) whether the English word 'white' is a word for white things. When playing this role, there can be no legitimately felt need to make the *T*-sentences come out true. Both roles are intelligible. But they cannot coherently be played both at once! The putative skeptical doubt is an absurdity in the English speaker's mouth; the Dutchman's doubt is intelligible—not skeptical, however, but empirical.

Why are the *T*-sentences true? They are sentences of our own language, and undeniable to us, though what they say is not necessary.

They are "pragmatic tautologies." In uttering them assertively we do no more than acknowledge that the quoted words all belong to our language and are properly put together. The word 'true' is a word of our own making, and we are carried along by its peculiar logic. No truth makers are needed to make the *T*-sentences true.

IX

Truth is not easy to come by. In this matter I disagree with Elgin, and agree—verbally—with Lewis. When I say this, however, I take myself to be paying due respect to the logic of 'true'; no more. Is truth on some interpretation at least easy to come by? Sometimes it is and sometimes not; but that is a different matter.[9]

But my agreement with Elgin runs deeper than any disagreement. Specifically, we share the conviction, expressed at the end of her paper and here in my previous section, that the issues under scrutiny belong to pragmatics, not to its impoverished derivative, pure semantics. Truth on some interpretation—that is prime subject matter for semantics, by definition. But truth is not exhausted there. When I use the word 'true' I am interpreting—hence, automatically and trivially, as Davidson points out, interpreting into my own language, the language I am using—and evaluating the accuracy of the text I am interpreting. This practice cannot be adequately dealt with if equated with a subcase of translation between languages, understood in a sense to which the question whether either or both are foreign or my own is irrelevant.

> One of the lessons of contemporary philosophy is that form is not detachable from content. My discussion here suggests that pragmatic considerations are not detachable either (302).

I agree.

BAS C. VAN FRAASSEN

Princeton University

---

[9] Think of the amount of effort that has gone into finding interpretations of scientific theories under which they may or might be true. Even in the case of Newton's mechanics and system of the world, let alone for quantum mechanics, the task has not been completed to everyone's satisfaction. But here I am speaking of genuinely admissible interpretations, by criteria that come to light as we progress. I am not speaking about the existence of functions of a certain type, which preoccupies ontology as done among the analytic philosophers.

BOOK REVIEWS

Reason and Rhetoric in the Philosophy of Hobbes. QUENTIN SKINNER. New York: Cambridge University Press, 1996. xvi + 477 p. Cloth $49.95.

A prospective reader of this book, turning to the table of contents, might be forgiven a moment of surprise. For the table shows the book divided into two almost equal parts, the first of which, in five chapters, is titled "Classical Eloquence in Renaissance England." Now, I have no reason to doubt that in these chapters, Quentin Skinner offers us a masterful survey, in depth, of this subject. But someone drawn to the book by a concern with the philosophy of Thomas Hobbes might wonder whether it tells her a good deal more than she needs, or perhaps even wants, to know. Such a reader might think that a single chapter on rhetoric and eloquence in sixteenth- and seventeenth- century England would have sufficed to acquaint her with the context of Hobbes's own work. Agreeing with Skinner that attention to the context of Hobbes's approach to rhetoric is of some importance, she might yet be reluctant to invest the degree of attention required by the extensive discussion in Skinner's book—especially as she comes to realize that this discussion contains few references to Hobbes, and those few very often take the form, "as Hobbes's translation [of Aristotle or of Thucydides] puts it," or words equivalent. Should she then pass from the introduction directly to the second part?

The second part, also in five chapters, is firmly on Hobbes—the title is "Hobbes and the Idea of a Civil Science." What is its argument? I think the main line may be put quite simply. Hobbes began as a humanist writer who accepted, with most of his contemporaries, the importance of the art of rhetoric. But he came to reject the cornerstones of rhetorical practice—the attempt to "start by securing so far as possible the goodwill and favourable attention of our audience" (257), the attempt to elicit "wise guidance in practical affairs...from a study of history" (259-60), and the willingness to accommodate "arguments to generally accepted beliefs and their expression in common speech" (263). He attacked the techniques of persuasion as unsuited to, and indeed destructive of, genuine teaching. He attacked also the ideal of the active citizen, as undermining political stability, and defended instead the obedient subject. For Hobbes, politics must be a science, and that science he set out, austerely, in *The Elements of Law* and *De Cive*.

Then Hobbes changed his mind again. Skinner quotes from *A Review and Conclusion* to *Leviathan*: "In all Deliberations, and in all

0022-362X/97/9402/94–97

Pleadings, the faculty of solid Reasoning is necessary: for without it the Resolutions of men are rash, and their Sentences unjust: and yet if there be not powerfull Eloquence, which procureth attention and Consent, the effect of Reason will be little" (352). The austerity of *De Cive* brought Hobbes few adherents; in *Leviathan* he tries to persuade as well as to prove (it brought him no more adherents). But a marriage of reason and eloquence cannot be an easy one; for, as Hobbes says, continuing the passage Skinner quotes: "But these [reason and eloquence] are contrary Faculties; the former being grounded upon principles of Truth; the other upon Opinions already received, true, or false; and upon the Passions and Interests of men, which are different and mutable."

In *The Elements of Law*, Skinner argues, Hobbes allied the moral to the natural sciences, insisting that the methods of the latter, directed at truth, were equally and solely appropriate to the former. But in his later writings, Hobbes rejects this alliance, continuing to insist that rhetoric has no place in the natural sciences, but eventually coming to accept it as necessary in politics. Indeed, Skinner claims that Hobbes comes to accept a threefold distinction between natural and moral science: (1) the former allows hypothesis, or supposition, whereas the latter proceeds strictly from definition; (2) the former are purely mechanistic, whereas the latter must "consider the purposes for which they [the artificial bodies investigated by civil scientists] are brought into existence"; (3) reason and eloquence may not be combined in the former, whereas in the latter they "may stand very well together" (355-56).

Skinner's claim here needs some amendment and comment. The first distinction to which he calls attention, between those sciences which do, and those which do not admit hypothesis, does not (for Hobbes) divide the natural from the moral sciences, but rather divides geometry and politics, both of which proceed strictly from definition, from all other sciences—the reason being, Hobbes insists, that in (and only in) geometry and politics, we produce by our own acts the objects about which we reason, and demonstration proceeds from generative definitions setting out these acts of production. The second distinction, which introduces purpose or (in Hobbes's later account in the *Decameron Physiologicum*) final cause to civil science, is never seriously examined by Hobbes, so that he does not explicitly consider the differences between demonstration in geometry and demonstration in civil science which it entails. It remains, therefore, largely latent in his work.

We may agree that these two distinctions show that Hobbes accepted significant methodological differences within the sciences, even if he failed to recognize fully some of their implications. But what of the third distinction? Is it methodologically significant? Does it affect "the central aspiration of Hobbes's civil philosophy, the aspiration to convert the study of moral and political theory into a scientific discipline" (1). I do not think that Hobbes believed that it did. I think it affected the way in which Hobbes found it appropriate to present different sciences, but not, in his view, the characteristics of the sciences presented. Only the first two distinctions, which do not bear on the role of rhetoric, affect those characteristics.

So let us turn to Skinner's own question: "What new insights into Hobbes's civil philosophy does it [Skinner's approach] yield" (8)? First, Skinner claims, he is "able to show how far Hobbes's conception of *scientia civilis* was shaped at all times by his changing reactions to the assumptions and vocabulary of classical and neo-Ciceronian theories of eloquence" (8). Certainly, Skinner is right to claim that Hobbes wanted to challenge the contemporary rhetorical understanding of civil science. But does this illuminate what Hobbes wanted to put in its place? Or do we understand that best by adverting to Hobbes's views of geometry and the natural sciences? I think the latter.

Skinner also claims that he is "able to establish a particular view of his [Hobbes's] intellectual development" (11). This is true, but what Skinner shows about Hobbes's intellectual development primarily concerns, if I am right, the differing ways in which Hobbes wanted to present his civil science, and not the character of that science itself. Hobbes's views on science did develop (the first two distinctions discussed above were not part of Hobbes's original position), but his changing stance on the role of rhetoric does not seem to me part of this development.

Finally, Skinner claims that he is "able to provide a new analysis of the rhetorical characteristics of *Leviathan*" (13). This is true. But does this analysis shed light on the argumentative changes between *Leviathan* and *De Cive*—changes which lead to the introduction into *Leviathan* of the problem of the Foole, the need for authorization, or the chapter on the rights of subjects, to take three important examples? I think not. The rhetorical character of *Leviathan* has misled some commentators into treating *De Cive* as "superior to it as philosophy,"[1] whereas (as I should be prepared to argue at length in an appropriate place) *Leviathan* is a significantly more sophisticated philosophical and scientific work than its predecessors. Skinner does not, of course, fall into the trap of downgrading the philosophy of *Leviathan* because of its rhetorical character. But he does fail to convince me, at least, that understanding its rhetorical character should affect our philosophical appreciation of the work.

---

[1] The quotation is from Bernard Gert's introduction to *Man and Citizen* (Garden City, NY: Doubleday, 1972), p. 3.

Skinner hopes that he "may have succeeded in raising anew the question of which style [rhetorical and dialogic, or rational and monologic] is more deserving of our intellectual allegiances" (16). Hobbes reluctantly accepted the need to persuade, but he never saw it as substitute for demonstration. We may understandably be less optimistic than Hobbes about the prospects of attaining demonstrative certainty in civil science, but I see no grounds for us to transfer our philosophical allegiance to the art of eloquence.

DAVID GAUTHIER
University of Pittsburgh

Kinds of Minds: Toward an Understanding of Consciousness. DANIEL C. DENNETT. New York: Basic Books, 1996. 184 p. Cloth $20.00.

When the old blind monk, St. Mael, had inadvertently baptized the penguins of Penguin Island, mistaking them for men, there was—so Anatole France[1] relates—a considerable stir in heaven. The Lord himself was embarrassed. He gathered an assembly of clerics and doctors, and asked them for an opinion on the delicate question of whether the birds must now be given souls. It was a matter of more than theoretical importance. "The Christian state," St. Cornelius observed, "is not without serious inconveniences for a penguin.... The habits of birds are, in many points, contrary to the commandments of the Church." After lengthy discussion, the learned fathers settled on a compromise. The baptized penguins were indeed to be granted souls—but, on St. Catherine's recommendation, their souls were to be of small size.

Daniel C. Dennett's new book is about the natural history of the souls—or intentional systems, as he would call them—that have evolved on earth. He begins with the very smallest, such as those of DNA, and works his way up to very largest, such as those of modern human beings, telling the story of their evolutionary relationships. It is a story of unashamed progress from lower to higher forms of men-

---

[1] *Penguin Island*, A. W. Evans, trans. (London: Watts, 1908/1931).

0022-362X/97/9402/97–103

tal life, in which large souls have amassed their powers by cunningly incorporating armies of relatively small ones. Dennett makes quite clear his belief that, if we human beings now see further than our distant ancestors, it is because we stand on the shoulders of dwarfs.

It is a book for laymen, attractively and simply written, taking little for granted, and avoiding much factual detail. But it is also a book that deserves close attention from those already familiar with Dennett's work. Dennett[2] has remarked elsewhere how his mentors, Gilbert Ryle and W. V. Quine, showed him "the importance of addressing an audience of non-philosophers even when they know that philosophers will be perhaps 95% of the actual and sought-for audience" (*ibid.*, p. 243). But, here, he shows how the converse can be true as well. For, while writing in this case primarily for nonphilosophers, he has kept the argument at a level that is clearly meant to interest and provoke his philosophical colleagues, too. Although he covers some familiar ground, he continually brings fresh perspectives to it: taking the opportunity to float new ideas, to revise some of his earlier ones, and generally to advance his campaign against those reactionary forces which still insist on there being an explanatory gap between the physical and mental worlds.

The most surprising development comes in his discussion of "the body and its minds." The idea that a person's body should be considered an integral part of his mind was latent in some of Dennett's earlier writings. But it seems to have been his reading of the neurologist Antonio Damasio's book, *Descartes' Error*,[3] that has now given him the confidence to make the idea fly. What Dennett now proposes is that the mind is distributed throughout the body. "One cannot tear me apart from my body leaving a nice clean edge, as philosophers have often supposed [including the author himself in his famous parable "Where am I?"]. My body contains as much of *me*, the values and talents and memories and dispositions that make me who I am, as my nervous system does" (77).

He is saying, in effect, that my mind needs my body not merely, as has always been recognized, as a means of gathering information and taking action in relation to the external world, but internally also—as a consultant, a judge, a resolver of disputes between otherwise unvalenced ideas. When confronted by a problem, my mind questions itself about what will work or make best sense, and begins the potentially vast search through a space of competing answers. But before you can say "combinatorial explosion!" my body on the basis of its evolved know-how and worldy expertise has already returned its own wise judgments. In due course, these judgments may be experienced by my conscious self as intuitions or gut feelings of pleasure or unpleasure, but probably not until well after they have already been decisive in directing my mind's business. The body, in short, is playing a role somewhere between Sigmund Freud's "id," a New Age angel, and a Prime Minister's "kitchen cabinet": and life, quite literally, would be unthinkable without it.

---

[2] "Dennett, Daniel C.," in Samuel Guttenplan, ed., *A Companion to the Philosophy of Mind* (Cambridge: Blackwell, 1994), pp. 236-44.

[3] New York: Putnam, 1994.

I expect anyone who reads of this idea—in Damasio's book or in the much clearer version here—will be struck by how obviously right it is. They will find they know in their bones that this idea of "knowing in one's bones" is a very good idea. But it may be even better than it immediately seems. For, as Dennett shows, it can be extended—fractally—in both directions: in from the body into the brain, and out from it into the external world. On one side, there are, as it were, mini wise bodies within the main body and mini-mini wise bodies within those. On the other side, there are mega bodies outside it in the culture and mega-mega bodies outside those.

Going inward, to start with, Dennett is set to wondering whether this stored bodily wisdom is to be found not just at the level of the sense organs or the muscles or the gut but in the elements of the brain itself. Indeed, maybe every individual nerve cell with its associated synaptic junctions should be considered to be in its own way a wise little bit of the mind: not simply a passive gate for information transfer but an embodied little intentional system in its own right—with its own beliefs, desires, and rationality. Only someone who, as Dennett does, thinks of all intentionality as being derivative rather than intrinsic, and believes in a version of what William Lycan has called ***homuncular functionalism***, would want to talk this way. But early in the book he has explained these views, especially in the excellent chapter on intentionality, where he argues as plainly as can be that the whole basis for the complex intentionality of higher minds is the delegation of work to relatively simple subminds. "Big minds"—so to speak—"have little minds within themselves to light 'em, and little minds have lesser minds, and so *ad infinitum*" (except that, since the minds do get simpler and simpler, this regress can eventually stop).

Dennett, as a homuncular functionalist, remains still committed to the general philosophy of functionalism. But he now becomes the first to acknowledge that, once we appreciate that the homunculi are not merely subminds but subbodies, too, classical functionalism is in some kind of trouble. It has been a central tenet of functionalism that, if mental states are the same thing as functional states, it does not matter what physical system is responsible for realizing them.

Hence all the thought experiments about minds made of beer cans, or Chinamen, or silicon chips. Even so, it has always been recognized that there are limits to such multiple realizability. While the mind's central processing can, in principle, be done by anything, the information gathering and the acting clearly cannot. You cannot, for example, make an *eye to see with* out of beer cans because beer cans are not photoreceptive, or a *mouth to kiss with* out of the Chinese population because the Chinese population is not wet. So, everyone agrees that functionalism has to yield to physicalism at the body's boundary, where the receptor and effector organs are. The new point, however, is that, if the brain is made up of body-like subunits, the same argument applies all the way in.

Thus, every nerve cell must be able to recognize particular patterns of electrical stimulation, chemical transmitters, and circulating hormones: it must "know" about morphine, or serotonin, or acetylcholine, for example, and it must adjust its own electrical and/or chemical output accordingly. But, in this case, you cannot make even a single nerve cell out of silicon chips, because silicon does not have the right sort of biochemistry. Let alone, can you make a full-blown human brain that can function in conjunction with a real human body.

As Dennett hastens to say, these considerations do not mean that functionalism is wrong as a metaphysical hypothesis. But they help make it clear that it is just that: a metaphysical hypothesis, and not a recipe—even in principle—for actually engineering a natural mind.

That is going inward. But going outward, Dennett joins up with another set of his old preoccupations. He has long been interested in how the use of well-designed tools and instruments can help make people smarter: a man with a pair of scissors, for example, is not just handier, he is in effect brainier—because he can now exploit his brain power in new ways. (I remember hearing Richard Gregory, to whom Dennett attributes this insight, wondering aloud about how much more of a genius he himself would have been if he had not been born fifty years before the invention of the personal computer.) But, by far, the most remarkable example of this trick of "importing intelligence from the cultural environment" is the use that human beings have learned to make of the mind tools of language.

So Dennett comes with gusto to a favorite theme: how the use of language has raised the minds of human beings to a new and superior order of reflection and inventiveness compared to any nonlinguistic animal. Now, however, he has a new way of talking about what language does for the human mind (although, surprisingly, he leaves it to his readers to draw the metaphor out). It is that language provides, in effect, a kind of culturally evolved body, a "body of experience and knowledge" if you like, which can function to speed up the mind's business in much the same way the corporeal body does. Human languages are rich in acquired know-how, common sense, manners, and mores. Hence, the body of language, just like the body of flesh and blood, can act as a sounding board against which to test the mind's evolving answers to its problems: expressing preferences, arbitrating between possible alternatives, and generally helping to bring matters to a quick and sensible conclusion. Language is not just a medium *in which* we think, it actually *does* some of the thinking with us and for us. Nor, in Dennett's hands, is this merely an "airy fairy" notion, that tells us little about human mentality in practice. As just one example of its fruitfulness, he goes on to make some quite specific—and so far as I know entirely original—suggestions about the way in which heard words may begin to form patterns in a child's mind before they have acquired any external referent, and in so doing prestructure the conceptual world.

Dennett is returning to his roots as he develops these themes. He prefigured many of the present arguments about the role of language in *Content and Consciousness*[4]—and he has never been one to abandon good ideas just because they have in the meantime become relatively unfashionable. Indeed, I detect that the air is loud with the sound of chickens coming home to roost. "My way of asking questions," he writes in the "Preface," "has a pretty good track record over the years.... Other philosophers have offered rival ways of asking the questions about minds, but the most influential of these ways, in spite of their initial attractiveness, lead to self-contradictions, quandaries, or blank walls of mystery" (vii). There will no doubt be readers—including followers of those other philosophers—who bridle at being thus warned in advance that they had better agree with Dennett or join the losers. But, though he says it himself, he *has* been right about many things, and on the evidence of this book he *will have been* right about a good many more.

A good many, but not all, I think. The strengths of the book are manifold. The weakness lies in the one area where Dennett has always seemed weak even to some of his friends, namely, in his cavalier treatment of the problems of sentience and qualia. In the present book he does not, as he has done sometimes in the past, simply push these problems to one side. In fact, he raises the issue of sentience early on and then returns to it again and again—like a Quiner returning to the scene of his crime. He asks: Is there a critically important difference between mere sensitivity and sentience? Is there a missing link—"an *x* factor"—which makes sensory experiences conscious? Is there a threshold, on one side of which lies nonconscious information processing, on the other side phenomenally rich experience?

---

[4] New York: Routledge, 1969.

These are for Dennett rhetorical questions, to which he is sure the answer must be "no." The fact that most sensible people would still imagine the answer to all of them is "yes" (think, for example, of the difference between painful stimuli processed during sleep and waking consciousness) does not faze him in the least. He is determined to scoff at the sentientialists. Among the several devices he uses to help him in this bad cause, I reckon the least justifiable is the appeal he now makes to moral sentiment. Our moral concern, Dennett suggests, should extend only to consciously sentient organisms. But everyone will agree, he goes on, that there is not and ought not to be a threshold of moral concern—an absolute dividing line between organisms that do and do not merit it. Hence, he implies, there cannot be any threshold for sentience.

As an argument this simply will not wash; and I am not sure Dennett himself really expects it to. But the fact that he argues this way at all only illustrates, I think, his frustration at not being able to get through and make his case in other ways. The pity is that he does not see that he may be making a strategic as well as a philosophical mistake in pursuing this campaign. For I would say that Dennett, of all philosophers, really has least need to be so embarrassed by the possible existence of an "*x* factor." Indeed, he could and should be leading the inquiry into what makes the difference between sensitivity and sentience, rather than obfuscating it. As we have seen, in other areas Dennett has been one of the first to argue that there have been qualitative leaps in the evolution of the mind. He himself has provided some of the best reasons for supposing that the involvement of language, for example, has made all the difference to human intelligence. Then why should not his own next move be to argue that the involvement of *the body* has made all the difference to sentience? I myself have suggested one scheme for how this could have come about.[3] Dennett says in his endnotes that he has not yet come to terms with my proposals. But I think that, without loss of philosophical face, he could come to terms with them—and no doubt, in doing so, improve on them. Indeed, I look forward to the day.

---

[3] *A History of the Mind* (New York: Simon & Schuster, 1992).

Dennett has not been right about this yet. But, maybe, after all, this is one more thing about which he *will have been* right when the day comes. I am reminded of an illusion that Dennett himself has helped to make famous, color "phi." If a red light goes out and a green light comes on a moment later a short way to one side, then when conditions are right an observer will see a light apparently moving smoothly across the space between and turning from red to green half-way across. The paradox is, of course, that it is as if the light somehow knows to turn green even before the green light has actually come on. Perhaps what we are seeing in this book is the spectacle of the world's most original philosopher of mind already beginning to change color, half way toward a still-to-be-formulated new theory of sentience, yet so far not ready to acknowledge it even to himself.

NICHOLAS HUMPHREY

New School for Social Research

A Realist Conception of Truth. WILLIAM P. ALSTON. Ithaca, NY: Cornell University Press. xii + 274 p. Cloth $35.00.

William Alston defends *alethic realism*, which he defines as the conjunction of two theses: (a) the realist conception of truth is the correct account of what it is for a proposition, statement, or belief to be true; (b) truth is important. The realist conception of truth is in turn summarized by a substitutionally quantified biconditional (28):

RCT (p) The proposition that p is true *iff* p.

Alston is willing to go beyond RCT to a minimalist correspondence account (38):

MCA (p) If the proposition that p is true it is made true by the fact that p.

He gives no account of facts, except to insist that they need not be mere shadows of propositions. He explains the application of 'true' to sentences as derivative from its application to the propositions that they express.

Alston usefully distinguishes between the concept and the property of truth. He intends RCT and MCA as accounts of the ordinary concept of truth, not of the property that it picks out. He allows that another concept of truth might pick out the same property by specifying its hidden essence; for example, the concept might embody a

0022-362X/97/9402/103–06

rich theory of facts. In this sense, Alston is a minimalist about the concept of truth, but not about the property of truth.

Does reference to the property of truth have explanatory value? In defending the importance of truth, the second conjunct of alethic realism, Alston uses considerations like this: when I act on a belief, I am more likely to achieve my goal if the belief is true. Minimalists about the property of truth object that the reference to truth is redundant, for it would have sufficed to say: when I act on a belief that *p*, I am more likely to achieve my goal if *p*. Alston replies that the disappearance of truth is only apparent. He claims that the generalizations in which 'true' does not occur are synonymous with generalizations in which it does, so the former employ the concept of truth just as the latter do (238). Thus, (1) must be equated in meaning with (2):

(1) (*p*) When I act on a belief that *p*, I am more likely to succeed if *p*.
(2) (*p*) When I act on a belief that *p*, I am more likely to succeed if that *p* is true.

The quantifier '(*p*)' in (1) is substitutional, so '(*p*)' had better be substitutional in (2), too. The phrase 'that *p*' in (2) refers to a proposition; otherwise, 'is true' does not have the sense in which Alston is interested. Presumably, (1) and (2) are synonymous only if their corresponding substitution instances are synonymous, and those substitution instances are synonymous only if their corresponding constituents are synonymous. Thus, (1) and (2) are synonymous only if the result of substituting 'Caesar was murdered' for '*p*' in the matrix of (1) is synonymous with the result of making the same substitution in (2), which is so only if the result of substituting 'Caesar was murdered' for '*p*' in '*p*' is synonymous with the result of making the same substitution in 'that *p* is true', that is, only if 'Caesar was murdered' is synonymous with 'That Caesar was murdered is true'. But elsewhere in the book Alston denies that 'Caesar was murdered' has the same content as 'The proposition that Caesar was murdered', on the grounds that only the latter is about a proposition (47). By the same token, 'Caesar was murdered' is not synonymous with 'That Caesar was murdered is true'; if they were synonymous, truth would not be a genuine property, which Alston insists that it is. Thus, (1) is not synonymous with (2). Although they are no doubt logically equivalent, that is not enough for Alston's defense of the importance of truth. That (1) is logically equivalent to something that employs the concept of truth does not entail that (1) itself employs the concept of truth, any more than the fact that (1) is logically equivalent to something that employs the concept of butter entails that (1) itself employs the concept of butter.[1] No alternative to (2) would do better for Alston; if any sentence in which truth is explicitly predicated of a proposition is synonymous with (1), (2) is.[2] Thus, Alston's defenses of the two conjuncts of alethic realism seem to conflict with each other.

Truth is important in at least this sense: there is an important difference between cases of true belief and cases of false belief. But it does not follow that we can distinguish between the two kinds of case only by referring to the property of truth. There is an important difference between cases in which the set of one's friends is the null set and cases in which it is not, but we can distinguish between the two kinds of case without referring to the null set.

Much of the book is a response to Brand Blanshard, Michael Dummett, and Hilary Putnam, who proposed epistemic accounts of truth. Alston has many effective points to make. The main focus of his attack is on epistemic accounts of the ordinary concept of truth. Some verificationists might accept RCT as an account of the ordinary concept, but insist that for some epistemic concept *E* (for example, justifiability in ideal epistemic conditions) we can know a priori that the ordinary concept and *E* pick out the same property. Alston argues much more tentatively against such epistemic accounts of the property of truth, suggesting that some highly theoretical propositions may be true without falling under *E* (203). Surprisingly, he does not consider examples used by Frederic Fitch, Bill Hart, and others. Let *p* be the proposition expressed by the sentence 'The number of hairs now on my head is even and no one will ever be justified in believing that the number of hairs now on my head is even', and *q* the corresponding proposition when 'odd' replaces 'even'. The number is either even or odd. Since no one will in fact bother to count, no one will ever be justified in having a belief either way. Thus, either *p* is true or *q* is. But no one could be justified in believing either proposition, even in ideal epistemic conditions: if someone were, the second conjunct would be false, so the whole proposition would be, contrary to the epistemic account of the property of truth.[5]

---

[1] If a substitutionally universal generalization is synonymous with the infinite conjunction of its instances, then (1) employs the concept of truth, because some of its instances do. For the same reason, (1) employs the concept of butter; such considerations do not show that truth is more important than butter.

[2] In Alston's own example, he claims synonymy between (3) 'For any *p* if *p* is asserted in the book, *p*' and (4) 'Everything asserted in the book is true'. (3) is ill formed, since '*p*' occurs in both name and sentence position; presumably '*p* is asserted' should be 'that *p* is asserted' and 'For any *p*' is a substitutional quantifier. Alston does not say whether the quantification in (4) is substitutional, but a difference in type of quantification would be hard to reconcile with the claimed synonymy.

There are some lapses. Alston argues that Dummett's verification-conditional semantics cannot use defeasible verification conditions:

> the statement of a set of confirming conditions, C, for a putatively non-conclusively verifiable sentence, S, is itself conclusively verifiable. But by the terms of verificationist semantics the statement of C is synonymous with S, since C is what specifies S's meaning. Therefore S is conclusively verifiable after all. Hence there can be no merely defeasibly verifiable meaningful sentences on this semantics. Q.E.D. (127).

This confuses specifying the meaning of *S* with providing a synonym of *S*. The verificationist semantics might entail something of this form about *S*, where '*c*' states *C*:

VS *S* is confirmed when and only when *c*.

Since Dummettian verificationists suppose verification conditions to be epistemically unproblematic, they will grant that when *c*, it is conclusively verifiable that *c*, as the first quoted sentence says. Since they will also grant that competent speakers of the language implicitly know VS, it is conclusively verifiable that *S* is confirmed when and only when it is conclusively verifiable that *c*. Putting these pieces together with VS, one concludes that when *S* is confirmed, it is conclusively verifiable that *S* is confirmed. What one cannot conclude is that when *S* is confirmed, *S* itself is conclusively verifiable. VS does not make *S* and '*c*' synonymous. For example, it does not imply that they have the same disconfirmation conditions.

I have concentrated on points of disagreement. Much in the book deserves agreement and applause; it is argued with care, subtlety, and good sense. If the ordinary concept of truth is not a very exciting one, that is not Alston's fault; it is just part of what he is telling us.

TIMOTHY WILLIAMSON

University of Edinburgh

---

[5] Such propositions are counterexamples to Alston's claim that 'The proposition that *p* would be known by an omniscient being *iff p*' is analytically true (53), unless there actually is an omniscient being.

NEW BOOKS: ANTHOLOGIES

BORCHERT, DONALD M., ed.: *The Encyclopedia of Philosophy Supplement.* New York: Macmillan, 1996. xxxii, 775 p. Cloth $100.00.

CARRUTHERS, PETER, and PETER K. SMITH, eds.: *Theories of Theories of Mind.* New York: Cambridge, 1996. xv, 390 p. Cloth $54.95, paper $19.95.

COHEN, JOSHUA, and JOEL ROGERS: *Associations and Democracy: The Real Utopias Project, Volume I.* ERIK OLIN WRIGHT, ed. New York: Verso, 1996. xiv, 267 p. Paper $18.95.

CRISP, ROGER, ed.: *How Should One Live? Essays on the Virtues.* New York: Oxford, 1996. vi, 263 p. Cloth $35.00.

ELDRIDGE, RICHARD, ed.: *Beyond Representation: Philosophy and Poetic Imagination.* New York: Cambridge, 1996. xii, 306 p. Cloth $49.95.

ENGSTROM, STEPHEN, and JENNIFER WHITING, eds.: *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty.* New York: Cambridge, 1996. ix, 310 p. Cloth $54.95.

HAAKONSSEN, KNUD, ed.: *Enlightenment and Religion: Rational Dissent in Eighteenth Century Britain.* New York: Cambridge, 1996. xii, 352 p. Cloth $59.95.

JORDAN, JEFF, and DANIEL HOWARD-SNYDER, eds.: *Faith, Freedom, and Rationality: Philosophy of Religion Today.* Lanham, MD: Rowman and Littlefield, 1996. xiv, 287 p. Cloth $58.50, paper $22.95.

JOSEPHSON, JOHN R., and SUSAN G. JOSEPHSON, eds.: *Abductive Inference: Computation, Philosophy, Technology.* New York: Cambridge, 1996. viii, 306 p. Cloth $44.95, paper $24.95.

KHALFA, JEAN, ed.: *What is Intelligence?* New York: Cambridge, 1996. vi, 207 p. Cloth $29.95, paper $19.95.

KREMER, ELMAR J., ed.: *Interpreting Arnauld.* Toronto: University Press, 1996. x, 183 p.

KVANVIG, JONATHAN, ed.: *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge.* Lanham, MD: Rowman and Littlefield, 1996. ix, 385 p. Cloth $62.50, paper $23.95.

LOVIBOND, SABINA and S.G. WILLIAMS, eds.: *Essays for David Wiggins: Identity, Truth and Value.* Cambridge: Blackwell, 1996. ix, 304 p. Cloth $54.95.

MASON, H.E., ed.: *Moral Dilemmas and Moral Theory.* New York: Oxford, 1996. xi, 246 p. Cloth $45.00.

MCCORMICK, PETER, ed.: *Starmaking: Realism, Anti-Realism, and Irrealism.* Cambridge: MIT, 1996. xix, 218 p. Cloth $30.00.

MORTON, ADAM, and STEPHEN P. STICH, eds.: *Benacerraf and His Critics.* Cambridge: Blackwell, 1996. xi, 271 p. Cloth $54.95.

MULHALL, STEPHEN, ed.: *The Cavell Reader.* Cambridge: Blackwell, 1996. vii, 420 p. Cloth $54.95.

PAPINEAU, DAVID, ed.: *The Philosophy of Science.* New York: Oxford, 1996. vi, 339 p. Cloth $55.00, paper $16.95.

PELTONEN, MARKKU, ed.: *The Cambridge Companion to Bacon.* New York: Cambridge, 1996. xv, 372 p. Cloth $54.95, paper $18.95.

NEW BOOKS: TRANSLATIONS

ALEXANDER OF APHRODISIASLEWIS: *On Aristotle's* Meterology 4. ERIC LEWIS, trans. Ithaca: Cornell, 1996. 186 p.Cloth $45.00.

AMMONIUS: *On Aristotle's* On Interpretation 1-8. DAVID BLANK, trans. Ithaca: Cornell, 1996. 206 p. Cloth $45.00.

ANSELM: *Monologion and Proslogion with the Replies of Gaunilo and Anselm.* THOMAS WILLIAMS, trans. Indianapolis: Hackett, 1996. xxi, 149 p. Cloth $34.95, paper $7.95.

AQUINAS, ST. THOMAS: *Commentary on the Book of Causes.* VINCENT A. GUAGLIARDO, CHARLES R. HESS, and RICHARD C. TAYLOR, trans. Washington, DC: Catholic UP, 1996. xxxvii, 193 p. Paper $16.95.

ARISTOTLE: *Politics: Book I and II.* TREVOR J. SAUNDERS, trans. New York: Oxford, 1995. xvi, 194 p. Cloth $55.00, paper $17.95.

ARNAULD, ANTOINE, and PIERRE NICOLE: *Logic or the Art of Thinking.* JILL VANCE BUROKER, ed. New York: Cambridge, 1996. xxxviii, 281 p. Cloth $54.95, paper $18.95.

BEISER, FREDERICK C., ed. and trans.: *The Early Political Writings of the German Romantics.* New York: Cambridge, 1996. xli, 203 p. Cloth $54.95, paper $18.95.

CASSIRER, ERNST: *The Philosophy of Symbolic Forms: Volume IV: The Metaphysics of Symbolic Forms.* JOHN MICHAEL KROIS and DONALD PHILLIP VERENE, eds. JOHN MICHAEL KROIS, trans. New Haven: Yale, 1996. xxvi, 237 p. Cloth $30.00.

CONWAY, ANNE: *The Principles of the Most Ancient and Modern Philosophy.* ALLISON P. COUDERT and TAYLOR CORSE, eds. and trans. New York: Cambridge, 1996. xxxix, 73 p. Cloth $49.95, paper $17.95.

DILTHEY, WILHELM: *Hermeneutics and the Study of History.* RUDOLF A. MAKKREEL and FRITHJOF RODI, eds. Princeton: University Press, 1996. xii, 409 p. Cloth $59.50.

GADAMER, HANS-GEORG: *The Enigma of Health: The Art of Healing in a Scientific Age.* Stanford: University Press, 1996. x, 180 p. Paper $14.95.

HAAR, MICHEL: *Nietzsche and Metaphysics.* MICHAEL GENDRE, ed. and trans. Albany: SUNY, 1996. xiii, 214 p. Cloth $49.50.

HEIDEGGER, MARTIN: *Holderlin's Hymn "The Ister."* WILLIAM MCNEILL and JULIA DAVIS, trans. Bloomington: University Press, 1996. xi, 185 p. Cloth $35.00.

LA METTRIE, JULIEN OFFRAY DE: *Machine Man and Other Writings.* ANN THOMPSON, trans. and ed. New York: Cambridge, 1996. xxx, 179 p. Cloth $54.95, paper $18.95.

MARGALIT, AVISHAI: *The Decent Society.* NAOMI GOLDBLUM, trans. Cambridge: Harvard, 1996. xi, 304 p. Cloth $35.00.

MCGUINNESS, BRIAN, and GEORG HENRIK VON WRIGHT, eds.: *Ludwig Wittgenstein: Cambridge Letters.* Cambridge: Blackwell, 1995. vii, 349 p. Cloth $49.95.

NANCY, JEAN-LUC: *The Muses.* PEGGY KAMUF, trans. Stanford: University Press, 1996. ix, 118 p. Paper $12.95.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 3, MARCH 1997

CALCUTTA UNIV. LIBRARY

THE RULE OF ADJUNCTION AND REASONABLE INFERENCE

One of the first rules of inference that every good logic student learns is the rule of adjunction: from P and Q, to infer P & Q. More appropriately, that from P and Q, it is *permissible* to infer P & Q. The rule can also be construed as formulating a doxastic obligation: if you believe P at t and believe Q, at t in the sense of full belief, you *must* believe P & Q at t in that same sense.[1] Or perhaps the point of the rule is that you are *committed* to believing the conjunction of any two things you believe. It is not clear what this notion of commitment comes to. I shall put my life on the line when it comes to the question of whether I shall be able to fly safely from Rochester to Cleveland; I shall do no such thing when it comes to the question of whether there will be an airline fatality within the next year. Nor is it all clear that doxastic obligation should be closed under conjunction: that I ought to do (believe) A and ought to do (believe) B does not entail that I ought to do (believe) A & B.

In some systems of logic, the rule of adjunction is even adopted as a primitive rule. In any event, it is a valid rule of inference: under any intepretation that makes the premises true, the conclusion will turn out to be true also. Or, put in terms of worlds: in any world in which P and Q are true, their conjunction P & Q will also be true.

Furthermore, in many systems of logic the rule of adjunction is a derived rule: it is derivable from whatever rules are taken as primitive. For example, if we have modus ponens as primitive, we can ar-

---

[1] It is a little difficult to construe logical rules, particularly the rule of adjunction, as representing obligations, since they so clearly cannot be fulfilled. One cannot even believe a single proposition A without failing to conform to adjunction construed as an obligation, since one would then be obliged to believe A & A, A & A & A,...and one would never get done.

0022-362X/97/9403/109–25

gue thus: taking the tautology $P \supset (Q \supset P \& Q)$ as a premise, two applications of modus ponens lead from the premises P and Q to the conclusion $P \& Q$, justifying adjunction as a derived rule.

I

So what can be the matter with the rule of adjunction? How can one possibly object to a valid rule of inference? Here is the other side. One object of logic is to distinguish what is permissible and what is not permissible in inference from a body of rational beliefs. (It could alternatively be claimed that it is rather to distinguish what is permissible and what is not permissible in inference from a body of *knowledge*; I shall consider that claim later.) If logic is construed permissively in this way, the rule of adjunction is in trouble on several counts; it is in worse trouble, of course, if we construe logic as formulating cognitive obligations.

First, we may consider the lottery paradox: if there is any probability such that any proposition having that probability should be accepted, then we can conceive of a lottery with so many tickets that, for any i, it should be accepted that ticket number i will not win. (We can easily factor in the probability that the lottery is dishonest.) But adjunction leads from 'Ticket 1 loses' and 'Ticket 2 loses' to 'Neither ticket 1 nor ticket 2 will win', and so, by another application of the rule of adjunction, to 'Neither ticket 1 nor ticket 2 nor ticket 3 will win', and so on. We are led to an explicit contradiction,[2] a self-contradictory statement: 'Some ticket wins and no ticket wins'.

Many writers (John Pollock,[3] for example) take arguments of this structure to be self-defeating, precisely because of the contradiction to which they lead. This reflects Pollock's principle of collective defeat. Others (Keith Lehrer,[4] for example) have devised ways to defeat the argument for a fair lottery, while accepting the claim that in an unfair lottery the most probable winner will win, even if the most probable winner will win with probability .00001: it suffices that every other candidate will win with probability strictly less than .00001). Nevertheless, the simplest procedure is to deny the reasonableness of adjunction in the case of either lottery, unless the unfair one has an assured (high probability) winner.

The fact that we are led to a contradiction is not the crucial issue: the same problem arises in the case of lotteries (like the New York

---

[2] As noticed in my *Probability and the Logic of Rational Belief* (Middletown, CT: Wesleyan, 1961).

[3] *Nomic Probability and the Foundations of Induction* (New York: Oxford, 1990).

[4] "Induction, Rational Acceptance, and Minimally Inconsistent Sets," in Grover Maxwell and Robert Anderson, eds., *Minnesota Studies in the Philosophy of Science*, VI (Minneapolis: Minnesota UP, 1975), pp. 295-323.

State Lottery) where the player gets to pick his own number: no winner is guaranteed, so no contradiction is entailed by the proposition that every player loses. But we all know perfectly well that someone will win, sooner or later.

Perhaps the lottery seems trivial or uninteresting. The same structure seems to raise its ugly head in default logic (as shown by David Poole,[8] for example). Similar considerations seem hard to avoid in scientific inference. Consider statistical tests designed to test hypotheses in medicine or psychology. One conventional structure for such a test is this: we test the null hypothesis H_0 of "no effect." In a well-defined sample space, we choose a region of rejection, such that, if the appropriate conditions of randomness, independence, and the like are satisfied, there is only a small chance (one in twenty, one in a hundred) that, on the supposition that H_0 is true, we shall get evidence falling in the rejection region. This is to say that, if H_0 is true, the probability of *falsely* rejecting it is less than that specified small chance. (We may also impose other constraints on the rejection region.) We do our sampling and obtain a sample in the rejection region. We reject the null hypothesis H_0 that the controlled variable has no effect. We do so on the statistical ground that we shall falsely reject H_0 when it is true with only (at most) that small preassigned chance.

The 'appropriate conditions' are not at issue here; we suppose that the conditions are satisfied and that we know that they are satisfied. The 'small chance' is not at issue here: what is small in one discipline might not be regarded as small in another. No matter; for the purposes of our discussion, merely consider a collection of inferences that *happen* to employ the same sense of 'small'—say, ϵ. The intuitive idea is that an experiment that leads to the rejection of the null hypothesis will be leading to error (that is, the null hypothesis will in fact be true, even though the evidence indicates otherwise) in a fraction no greater than ϵ of cases. Or, thinking positively, in general we will not be misled in a fraction of at least $1 - \epsilon$ of cases. We may therefore feel comfortable in supposing that the results of *this* experiment are not misleading.

But, of course, if we consider a collection N of experiments, we can be just as comfortable in believing that at least one of them *is* misleading. Now, this is not ordinarily troublesome, because we do not ordinarily make inferences based on a large number of these conclusions as premises. But the principle of conjunctively combining such conclusions certainly seems suspect, nevertheless: the chance that both null hypotheses are nonmisleading, in the case of testing two hypotheses, is $(1 - \epsilon)(1 - \epsilon)$, which is unarguably less than the chance that one of them is. If one imagined a scientific journal containing reports of N of these experiments, the chance that one of them was misleading could (by a correct choice of N) be made greater than $(1 - \epsilon)$.[6]

---

[8] "The Effect of Knowledge on Belief: Conditioning, Specificity, and the Lottery Paradox in Default Reasoning," *Artificial Intelligence*, XLIX (1991): 281-307.

Note, incidentally, that these high probabilities are, if our assumptions are satisfied, independent. Thus, Pollock's argument (see below) designed to block inference in the case of the lottery does not seem to apply here any more than it does to measurement: there are no grounds for collective defeat.

It is hard to believe that "inference to the best explanation"[7] does not also lead to conclusions that are conjunctively suspect. Suppose we have two explananda, E_1 and E_2. What reason have we to believe that, in general, the conjunction of the best explanation of E_1 and the best explanation of E_2 is the best explanation of E_1 & E_2? For example, the best explanation, given what we know, K, of the failure of the accounts to balance, E_1, may be that a bookkeeper made a mistake; the best explanation of Smith's trip to Brazil, E_2, given K, may be that he wants a winter vacation; but given this background knowledge K, the best explanation of E_1 & E_2 may be fraud.

There is also the example of measurement, fundamental to almost any scientific enterprise, to be considered. No procedure of measurement is perfect; all admit of error. What we can rationally conclude from a measurement is not that the magnitude m has such-and-such a real-number value, but that its value lies within d units of the result r of the measurement: $m \in (r \pm d)$. The justification of this conclusion is (like that of the conclusion that ticket i will not win) that it is simply incredible (too improbable to believe) that an error of magnitude greater than d has been committed (it is simply too improbable to believe that ticket i will win).

There is an argument, due to Pollock (*op. cit.*; it is also mentioned by Gilbert Harman[8]) that seeks to drive a wedge between the lottery and these other cases. It is based on independence: that measurement number 1 is in error by less than d does not increase the probability that measurement number 2 is in error by less than d—the errors of measurement are independent. That ticket number 1 loses decreases the probability that ticket number 2 will lose—the success of lottery tickets is not independent. Put another way, we could say that there is no contradiction in supposing that any number N of measurements are each in error by less than d, though, when N is the number of tickets in a conventional lottery, there is a contradiction in saying that it is a lottery and that all N tickets will lose.

---

[6] Assuming, of course, that it makes sense to talk about such chances. Most classical statisticians would balk at that way of talking, and insist that 'chance' only makes sense after you have specified a sample space and a measure on it, which we have not done. We are speaking here informally and intuitively.

[7] Gilbert Harman, "The Inference to the Best Explanation," *Philosophical Review*, LXXIV (1965): 88-95.

[8] Pollock takes the two cases to be distinguished to be the lottery paradox and the paradox of the preface. Harman, *Change in View* (Cambridge: MIT, 1989), p. 71.

The response is that we are not dealing with necessities, but with reasonable beliefs. The grounds on which we accept $m \in (r \pm d)$ are that the chance of making an error as large as d is suitably small. On the same statistical grounds, employing the same sense of 'small', we can say that the chance that of N measurements none will be in error by more than d is just as suitably small. This is a consequence of exactly that independence of errors that was alleged to reflect the difference between the lottery and the measurement. For large enough N, the conjunction of N statements of the form $m_i \in (r_i \pm d)$ will be almost certainly false. As Pollock notes with respect to the preface, there is no contradiction in supposing that all N statements of the book are true; similarly, there is no contradiction in supposing that all N measurement statements are true; but it is an incredibility that only the most naive would swallow.

Even the results of two measurements should not be conjoined, if they are measurements of the same distance D. Suppose we measure this distance, and conclude that its value is 3.35 ± .05. Suppose our laboratory mate measures it, and concludes that its value is 3.32 ± .05. If we were to conjoin these statements, we could conclude that its value is between $3.30 = 3.35 - .05$ and $3.37 = 3.32 + .05$. This is obviously quite unlikely. In reality, what we can conclude (supposing independence of the measurements) is that the value of D is $3.335 \pm (.05/\sqrt{2})$.

Another kind of argument that undermines the authority of the rule of adjunction is a general one. It does not take an excess of modesty to believe as firmly as one believes almost anything one believes that something one believes is false (leaving to one side mathematical and logical truths). If we accept the rule of adjunction, then to accept any set $(p_1,\ldots,p_n)$ of beliefs is to be entitled to accept their conjunction. But to believe that one of these statements is false is exactly to believe the denial of that conjunction.

There is no need here to put this claim in a metalanguage. Suppose that one believes (explicitly) a finite number of statements. To believe that one believes something false is exactly to believe in the denial of the conjunction of this set of statements. There is nothing mysterious either about the syntax or the epistemology. Of course, to *believe* the denial of the conjunction of (all of) what one believes is easy if the statement expressing this state of affairs is one of the conjuncts in question: matters could hardly be otherwise! The form of one's corpus becomes $p_1, p_2, \ldots, p_n, \neg(p_1 \,\&\, p_2 \,\&\ldots\&\, p_n)$, where the first n statements as premises yield $(p_1 \,\&\, p_2 \,\&\ldots\&\, p_n)$, and it is certain that one of these two statements is false.

But let us try to avoid this trivialization of the problem. Suppose we consider that there is a finite basis for our reasonable beliefs. These are statements in some distinguished sublanguage of our general language. They need not necessarily be "protocol" statements in the sense of the Vienna Circle, but some set of statements accepted on some grounds other than inference—for example, the statements obtained by Isaac Levi's[9] "routine expansion" of a corpus of knowledge, statements that are prima facie acceptable. If we accept these statements individually as reasonable beliefs, we are led by adjunction to accept their full conjunction as a reasonable belief. But if those statements have empirical content, they must be (perhaps only very slightly) uncertain. Their conjunction, then, must be even more uncertain. If we suppose no more than that the probability of the ith statement, given the truth of the preceding i-1 statements, is less than 1 by some finite amount greater than ϵ, then for large enough N, the probability of the conjunction of N such statements is itself less than ϵ. In the case of probabilities, whatever probabilistic support there is for the N individual statements may be exceeded by the probabilistic support for the denial of their conjunction.[10]

What is more to the point would be the denial that these beliefs are a reflection of probabilistic support at all. It is surely arguable that the epistemic justification for these basic beliefs is not merely probabilisitic. Nevertheless, few people would maintain that basic beliefs are incorrigible. The process that gives rise to them is nonmonotonic; even observational justification is nonmonotonic. Any such process admits of error, else we would not need to regard it as nonmonotonic—we could always retain an error-free set of beliefs. But if our basic beliefs admit the possibility of error, it is hard to see how this possibilty of error can avoid being magnified by the adjunction of many basic beliefs. At some point the possibility of error is too great to allow the acceptance of the conjunction. Adjunction would require that we accept the conjunction of these basic statements, however many there were. This seems unreasonable.

---

[9] *The Enterprise of Knowledge* (Cambridge: MIT, 1980).

[10] Neither the rule of adjunction, nor the arguments herein, bear on propositions of the form '*S* is probable'. Probability is mentioned only as a typical *ground* of full acceptance.

II

We must still consider the semantic justification of the rule of adjunction: if each of $p_1 \ldots p_n$ is true, then clearly their conjunction *must* be true. So how can there be anything wrong with the rule of adjunction? Surely, it will be said, if I *know* A and B, I am also entitled to claim that I *know* their conjunction.

Let us be careful. True, on most analyses of knowledge, if I know A and B, then by most standards I know their conjunction. It is presupposed that knowledge implies truth: if I know that A, and I know that B, both A and B are true, and therefore (of course) their conjunction is true.

But there is a gap between my *claim* that I know A, and the fact of my knowing A. It may be a small gap, but with regard to adjunction we should regard it as unbridgable. Although my knowledge claims, conjoined, may imply a claim regarding the truth of A & B, the claim that they imply will naturally in some cases be weaker than the original claims. We surely cannot plausibly suppose that a conjunction of claims will always be as strong as the individual claims.

There is no universally agreed upon analysis of knowledge; it is not clear that the most useful analysis will turn out to be one in which knowledge implies truth. If this is so, it may be that the best construal of knowledge will not have the semantic property that, if A and B are known, then their conjunction will also be known. This is an issue that should be considered quite independently of the popularity of adjunction as a principle of logic.

The upshot is that, if premises of empirical argument are known, and knowledge entails truth, then the rule of adjunction, being both knowledge preserving and truth preserving, is unexceptionable. But if the premises are *claims* to knowledge, then the conclusion of empirical argument—also a claim to knowledge—may constitute a weaker claim than is constituted by each of the original premises individually.

III

With the weight of skepticism mounting against adjunction, it is time to take a look at systems of logic that might allow us to deny it. The systems to look at are systems of paraconsistent logic—systems in which global inconsistencies of belief can be tolerated. There have been a number of studies of such systems.[11] In general, the goal is to handle the so-called semantic paradoxes: the paradox of the liar, for example, or to provide a way of handling 'true' as a predicate in the object language. What is denied, in the first place, is either the law of the excluded middle ($T \vee F$) or the law of contradiction $\neg(T \wedge F)$ or both. The former approach is followed by Donald Perlis,[12] and is, of course, the path pursued by multivalued logics (for example, the family of logics studied by Newton C. A. da Costa (*op. cit.*)). The latter is the approach followed by Graham Priest (*op. cit.*), who embraces not only inconsistent bodies of statements, but also explicit inconsistencies of the form '$\alpha \wedge \neg \alpha$'. Perlis, who favors an approach in which no well-formed formula can be both true and false, holds to a perfectly classical logic, including the principle of adjunction. The same is true of the logics studied by da Costa, who takes as a principle of valuation the classical principle that A is true in M and B is true in M if and only if A & B is true in M. Adjunction is therefore a valid rule of inference.

---

[11] There is a fairly complete history of paraconsistent logics in Graham Priest, Richard Routley, and Jean Norman, *Paraconsistent Logic: Essays on the Inconsistent* (München: Philosophia, 1989). Modern players include Priest (*In Contradiction* (Dordrecht: Reidel, 1987); "Reasoning about Truth," *Artificial Intelligence*, xxxix (1989): 231-44); Nicholas Rescher and Robert Brandom (*The Logic of Inconsistency*, Totowa, NJ: Rowman and Littlefield, 1979); Peter K. Schotch and R.E. Jennings ("Inference and Necessity," *Journal of Philosophical Logic*, ix (1980): 327-40; "Non-Kripkean Deontic

Priest favors an approach in which the law of contradiction is rejected, thus allowing certain sentences (for example, the sentence L of the liar paradox) to be both true and false. He must also then weaken his logic, so that it is no longer possible to derive all formulas from a contradiction. (*Ex contradictione quodlibet*, as we Latin scholars put it.) Nevertheless, since classical valuations are merely extended in his system to include the possibility of 'both true and false', the inference from A and B to A & B remains valid: if A and B are both assigned T, their conjunction will also be assigned T.

While Priest allows truth assignments to be chosen from among (true, false, true-and-false), and Perlis takes truth values to be chosen from among (true, false, neither-true-nor-false), Nicholas Rescher and Robert Brandom (*op. cit.*) seek greater generality by allowing

---

ed. *New Studies in Deontic Logic* (Dordrecht: Reidel, 1987), pp. 149-62; "On Detonating," in *Paraconsistent Logic: Essays on the Inconsistent*, pp. 302-327); Newton C.A. da Costa ("On the Theory of Inconsistent Formal Systems," *Notre Dame Journal of Formal Logic*, XV (1974): 497-510; da Costa, Lawrence J. Henschen, James J. Lu, and V.S. Subrahmanian, "Automatic Theorem Proving in Paraconsistent Logics: Theory and Implementation," *10th International Conference on Automated Deduction* (New York: Springer, 1990), pp. 72-86; A.I. Arruda, da Costa, and Rolando Chuaqui, eds., *Non-Classical Logics, Model Theory, and Computability* (Amsterdam: North-Holland, 1977)); and Eliezer L. Lozinskii ("Information and Evidence in Logic Systems," *Journal of Experimental and Theoretical Artificial Intelligence*, VI (1994): 163-93; "Resolving Contradictions: A Plausible Semantics for Inconsistent Systems," *Journal of Automated Reasoning*, XII (1994): 1-31).

[12] "Languages with Self-reference I: Foundations," *Artificial Intelligence*, XXV (1985): 25-79; "Languages with Self-reference II: Knowledge, Belief, and Modality," *Artificial Intelligence*, XXXIV (1988): 179-212; "Truth and Meaning," *Artificial Intelligence*, XXXIX (1989): 245-50.

truth values to be chosen from among (true, false, true-and-false, neither-true-nor-false). Like the authors previously mentioned, Rescher and Brandom are primarily concerned with the semantic paradoxes, and with finding intuitively appealing ways to avoid them. Nevertheless, unlike these other authors, they attempt to provide an intuitive background of operations on possible worlds that make sense of the logic they develop.

Let us take 1 for true, and 0 for false. Then the four truth values are (1, 0, (1,0), ∅). Every proposition in a standard world has one of the classical truth values 0 and 1. But we can obtain nonstandard worlds by *schematization* (taking a schematic world to be one having truth-value gaps—that is, one in which certain statements are neither true nor false) or *superposition* (taking an inconsistent world to be one having overdetermined truth values—that is, one in which certain statements are assigned both the value true and the value false).

The idea behind these operations on worlds is to reflect a syntactic operation: that of forming unions and intersections of sets of sentences. If S_1 and S_2 are sets of sentences, one may contain a sentence A and the other the negation of A. There is no standard world that models the union of S_1 and S_2, but there is a superposed world. This world will assign the pair of truth values (1,0) to A, but need not be abnormal in its assignment of truth values to any other sentence. Similarly, the intersection of S_1 and S_2 will correspond to a schematic world, in which sentences that do not appear in both S_1 and S_2 (for example, $\neg A$) are assigned the empty truth value ∅. On this scheme, an explicit contradiction, $A \wedge \neg A$, is always false—it always gets the truth value 0—in any world, standard or not. In this it is to be contrasted with the views of da Costa and Priest, according to which we can have reason to accept $A \wedge \neg A$—that is, we can have good reason to accept A and good reason to accept $\neg A$ (shades of G. W. F. Hegel and Karl Marx!).

The semantics of inference, and the logic that represents it, is perfectly standard for any standard world. Where it breaks down is exactly in connection with inconsistent worlds. It is only in inconsistent worlds that the principle of adjunction fails to hold.[13] For reasons given above, this is not enough. There is nothing inconsistent about the world in which we have 10^{10} measurements, each of which is accurate within two standard deviations of error; yet we should not want to conjoin the results of all those measurements; we can be dead certain that some of them are wrong, even though this certainty is not a logical truth.

---

[13] Rescher and Brandom, *op. cit.*

Peter K. Schotch and R. E. Jennings[14] provide a semantics for globally inconsistent bodies of knowledge which is more nearly in accord with the intuitions we have been trying to cultivate. They are concerned with practical questions—with bodies of statements we might take as representing reasonable beliefs. Their suggestion is that we should consider maximal consistent subsets of this body of statements.[15] We may then consider statements that are in every such subset; statements that are in some such subset; and so on. Within a maximal consistent subset of *S*, adjunction is fine, though it is not applicable to the corpus of statements as a whole. What is applicable to the corpus as a whole is a form of adjunction in which premises can be combined, for example, as follows: if the number of maximal consistent subsets of a set of statements is two, then we can combine three premises, *A*, *B*, and *C*, by adding the set of pairwise disjunctions of these statements, $(A \,\&\, B \vee A \,\&\, C \vee B \,\&\, C)$, to our argument, on the sensible ground that one of these conjuncts, at least, must appear among the (classical) deductive consequences of one of the maximal consistent subsets.[16]

Again, this result is too permissive. In the case of the lottery, a maximal consistent set of sentences describes in detail the way the lottery turns out: ticket #47943 wins, and all others lose. It is not reasonable to conjoin arbitrary subsets of such a set of sentences, even though they are consistent. As the examples offered above suggest, we should abjure all forms of adjunctive inference. The question is: What does this leave us? Is this preposterously limiting? Only if we have the idea that somehow premises should be "simple" or "basic." There is no need to adopt this idea. If we have reason to accept p_1 and to accept p_2 and...to accept p_n we may also have good reason also to accept their conjunction $p_1 \,\&\, p_2 \,\&...\&\, p_n$. (If we do not, their use as premises is surely suspect!) We need only deny that this is *always* the case, for any *p*s and any *n*, which is what the rule of adjunction requires of us.

To be cautioned about the acceptance of conjunctions of premises is not unreasonable. An argument (in real life) is often designed to justify the acceptance of some conclusion on the basis of premises that the arguee finds acceptable. Do you accept the premise p_1? Do you accept the premise p_2? Then here is an argument in the first-order predicate calculus that shows that p_1 and p_2 imply the conclusion Q. What our considerations are showing is that we should be cautious about proceeding from the acceptance of p_1 and the acceptance of p_2 to the acceptance of their conjunction. But that is something we knew all along: acceptability is not adjunctive. An argument that used enough empirical premises (say, a million), could surely use premises that are individually acceptable, but lead to a completely unacceptable conclusion. We have offered a number of examples of this.

---

[14] "On Detonating."

[15] These correspond to the *strands* of my *The Logical Foundations of Statistical Inference* (Dordrecht: Reidel, 1974); they are similar to the degrees of inconsistency of my "Conjunctivitis," in Marshall Swain, ed., *Induction, Acceptance, and Rational Belief* (Dordrecht: Reidel, 1970), pp. 55-82.

[16] Details are to be found in Schotch and Jennings, "On Detonating," and in P.J. Apostoli and Bryson Brown, "Compactness and Completeness for Generalized Modal Logics" (forthcoming).

What we need for a logical argument to be persuasive is not merely that each of the empirical premises be acceptable to us, but that their conjunction be acceptable. This is just to say that a persuasive argument proceeds from essentially one premise: an argument from a set of premises (P_i) to a conclusion Q is acceptable just in case there is some finite *acceptable* conjunctive premise $P = P_j \& P_k \& \ldots \& P_l$, where each conjunct is in the set (P_i), such that $P \supset Q$ is a theorem.

IV

What impact does this proposal have on traditional applications of logic? It is certainly true that many arguments involve multiple premises. Even humble modus ponens is usually presented as involving the inference from A and $A \supset B$ to B—clearly, a pair of premises. In fact, in many systems of logic[17] modus ponens is the only rule of logic. Can we get along instead with only the one-premise rule, from $A \& A \supset B$ to infer B? No, but we may make use of the fact that when A and B are logical truths, we do have as much warrant for their conjunction as we have for them individually. It is also true that when B is a logical truth, we have as much warrant for the conjunction of A and B as we have for A. Thus, if A is an empirical premise, and the conditional $A \supset B$ is a logical truth, we have as much warrant for $A \& A \supset B$ as we have for A itself. Thus, we have at least as much warrant for B as for A, and the argument based on modus ponens is indeed as good as the argument for the empirical premise.

Of course, we must distinguish this use of modus ponens from that in which both the conditional and the antecedent are merely empirically justified premises. In that case, the rule of adjunction is essential, and we must, in order to have an adequate justification for the conclusion, have adequate justification for the conjunction of the premises, and not merely for each premise separately. Here, of course, is the place to represent the reluctance of some philosophers to accept the gloss '$\neg(A \,\&\, \neg B)$' for the conditional. The evidence may support '$A \supset B$' mainly because it supports '$\neg A$'. The evidence may also support A; but the total evidence should not be taken as supporting B.

---

[17] For example, W.V. Quine, *Mathematical Logic* (Cambridge: Harvard, 1951); Herbert Enderton, *A Mathematical Introduction to Logic* (New York: Academic, 1972).

V

Two common forms of argument have been extensively discussed in connection with the possible inconsistency of our bodies of belief. One is reductio ad absurdum. Reductio has been the subject of suspicion for many years, and this suspicion is one of the driving forces, if not the main driving force, behind the development of many-valued logic. If our beliefs can be inconsistent anyway, what does it matter that an additional hypothetical belief leads to contradiction? What is the force of the argument of the form: suppose that A; then we are led to both C and $\neg C$; therefore, we must deny A: therefore, we can accept $\neg A$? Bryson Brown[18] finds fault with the system of Rescher and Brandom for the very reason that it cannot account for this natural kind of inference. He goes on to praise Schotch and Jennings for giving an account of it within the framework of inconsistent bodies of knowledge.

The account works like this: a set of statements S that is inconsistent may be represented as a set of sets of maximal consistent subsets of S. Thus, A, $A \supset B$, and $\neg B$ is an inconsistent set of statements. The set of maximal consistent subsets are $(A, A \supset B)$, $(A, \neg B)$, and $(A \supset B, \neg B)$. The minimal number of subsets whose union is equivalent to S is a measure of the level coherence of S. The idea of Schotch and Jennings is that, if we add a statement T to S, and the incoherence of S and (T) is greater than that of S, we should not have added T in the first place. This is a reductio proof of $\neg T$: "a sentence is absurd with respect to some premise set if and only if adding it to the set increases the set's level [of incoherence]" (*ibid.*, p. 8).

The approach of Schotch and Jennings, like that of classical logic, treats maximal consistent sets of sentences as conjunctions. Thus, to use the structure of the lottery as an example, each maximal consistent subset of the set of acceptable sentences will contain a complete description of the lottery: the lottery is fair; there are a million tickets; there is one winner; every ticket other than ticket *i* will lose; and (therefore) ticket *i* will win. As a conjunction, of course, the story embodied in this set of statements is as improbable as 'Ticket *i* will win' itself. There will be, of course, a million and one maximal consistent sets corresponding to the million ticket lottery (counting the assertion of fairness as also suspect); the level of incoherence of any body of knowledge that contains the lottery description will be at least a million and one. Adding one more contradiction does not seem to add much incoherence to what we already have.

---

[18] See "Rational Inconsistency and Reasoning," *Informal Logic*, XIV (1992): 5-10, esp. p. 8. The measure of incoherence adopted reflects the number of statements involved. This seems an odd measure in the sense that the bald contradiction (P, $\neg P$) exhibits a small level of incoherence (2), while the description of the lottery, with probabilistic acceptance, exhibits a level of incoherence of over a million.

This seems unnecessarily complicated. When we employ reductio, we do not simply poke around in our body of knowledge, looking for any old contradiction. We have in mind a context: the axioms of geometry or number theory; the description of the characters and circumstances in a detective story; or whatever. We have a particular *consistent* set of premises in mind, which we are willing to take as an acceptable conjunction. Now we can say that, if there is good reason to accept the conjunction of these premises, and this conjunction is logically inconsistent with $\neg T$, there is equal reason to accept T. Do we want to apply reductio when the context is represented by a set of consistent statements whose conjunction we are unwilling to accept? It does not seem to me that we should want to give more credence to the conclusion than we are entitled to give to the conjunction of those premises required for the reductio.

VI

A more recent worry, and one which has had less ramified effects in logic, is the worry about argument by cases.[19]

Simpson's paradox[20] provides one motivation for worrying about the argument by cases. It is perfectly possible that the frequency of a target property is increased by conditioning on a test property in each partition of a population, and at the same time that it is decreased by conditioning on the test property in the population as a whole.

For example, suppose we are concerned with the efficacy of a cold remedy. We conduct two trials: one on students, and one on faculty. The numbers are as follows:

---

[19] Ronald P. Loui, "The Case against Cases" (unpublished).

[20] Colin Blyth, "On Simpson's Paradox and Mutually Favorable Events," *Journal of the American Statistical Association*, LXVIII (1973), p. 746; Eric Neufeld, "Simpson's Paradox in Artificial Intelligence and in Real Life," *Computational Intelligence*, XI (1995): 1–10.

	Treated	Untreated
Students		
total	1000	100
improved	400	20
Faculty		
total	100	1000
improved	90	800

Table 1

The remedy seems efficacious for undergraduates, and it also seems efficacious for faculty; but for the university population as a whole it seems counterproductive: the fraction of those treated who improved is smaller than the fraction of the untreated group who improved!

While this example does not embody an argument (even a nonmonotonic argument) by cases, it is certainly suggestive. If X is a member of the university community, X must be (we are supposing) either a faculty member or a student. If he is a faculty member, the chances of improvement are greater if he takes the remedy. If he is a student, the chances of improvement are greater if he takes the remedy. But the evidence says that, if he is a student *or* a faculty member, the chances of improvement are *less* if he takes the remedy. Knowing he is either a student or a faculty member, but not knowing which, do we recommend that he take the remedy? Argument by cases says "yes"; direct inference says "no."

Ronald P. Loui notes that nonmonotonic logics do not support argument by cases (*op. cit.*). He raises the question of whether these logics should allow argument by cases, but does not offer a definitive answer. Sometimes our intuitions seem to support argument by cases, and sometimes not, and sometimes our intuitions are simply not clear.

Being careful about adjunction again holds the key. $(A \vee B)$ & $(A \supset C)$ & $(B \supset C) \supset C$ is simply a tautology: we cannot find fault with this. But this is just what does not persuasively translate into the language of support or the language of rational acceptance. Sometimes the disjunction of the cases is a logical truth; other times it is not. What is more important is that the conditionals are usually not logical entailments. It is not from A alone, but from A together with other stuff X_A that we have reason to believe, C follows, and from B, together with other stuff X_B that we have reason to believe, C follows. We *represent* these entailments by conditionals $A \supset C$ and $B \supset$

C, and use the tautology. But if we have reason to be suspicious of adjunction, we had better unpack these entailments: what we really have are $X_A \rightarrow (A \supset C)$ and $X_B \rightarrow (B \supset C)$, where the arrow represents logical derivability. Thus, the form of the disjunctive argument is $(A \vee B) \& X_A \& X_B \rightarrow C$. We need the conjunction of the antecedents. While this may often be forthcoming in an argument that depends on a small number of cases, it can become quite problematic in an argument that requires hundreds of cases, as Loui's examples show.

VII

Our conclusion is to be contrasted with that of Rescher and Brandom (*op. cit.*) as well as that of Schotch and Jennings.[22] Rescher and Brandom retain the deductive principle of adjunction (*op. cit.*, p. 18), and reject instead the semantic principle: that from the truth of A and the truth of B, you can infer the truth of $A \& B$. Their choice stems from two facts.

First, they wish to show that it is possible to accommodate conflicting assertions in the same system. Thus, they want to be able to regard both A and $\neg A$ as acceptable theses, though they wish to reject (as we would) the thesis $A \& \neg A$. Thus, $\{A, \neg A\}$ yields $A \& \neg A$, but the truth of $A \& \neg A$ does not follow from the truth of A and the truth of $\neg A$. Their interest in this possibility has two sources: it stems from an interest in finding a new way to deal with the logical paradoxes—Russell's paradox as well as semantic paradoxes, such as that of the liar; and it stems also from an interest in finding a way to make sense of the Hegelian and Marxist claims that contradiction is inherent in descriptions of the world.

Second, the means by which Rescher and Brandom seek to do this is through a possible-worlds analysis. The basic idea is that of combining possible worlds. If there is a world in which A is true and a world in which $\neg A$ is true, one way of combining those worlds will leave A with no truth value (underdetermination) and another way of combining them (superposition) will make both A and $\neg A$ true—though, of course, it will not make $A \& \neg A$ true.

The semantic rule of adjunction—that, if A is true in a world w and B is true in a world w, then $A \& B$ is true in that world—fails on their analysis only for inconsistent worlds. It fails for these worlds, as a principle, because in the inconsistent worlds both A and $\neg A$ may be true, and we do not want to infer that $A \& \neg A$ is true, since this is false, on logical grounds alone, in any world whatever.

---

[22] "On Detonating."

On the analysis offered earlier, it seems that the syntactical principle of adjunction should fail in many more cases than these. In particular, it will fail for many sets of sentences, each of which will be justifiable in its own right. We are trying to model inference for epistemology; Rescher and Brandom are modeling inference for logic and mathematics, and possibly for metaphysics. We are focusing on inference from premises; Rescher and Brandom are focusing on logical truth. On their construction, the classical logical truths still come out to be true in every world (to be implicit in every set of sentences).

The approach adopted by Schotch and Jennings is closer to that adopted here. It is syntactical rather than semantic, and the number of cells, if we consider lotteries or measurements, can be extremely large. A million ticket lottery, for example, gives rise to a million and one cells, since that is the number of distinct maximal consistent subsets of our assumptions. The most important difference from the present approach is that these consistent subsets are treated in the classical way: they are taken to be (potentially) closed under deduction. But, clearly, if we are expanding a subset of the statements comprising an inconsistent body of knowledge right up to the point where the next statement to be accepted will produce an inconsistency, the conjunction of the preceding statements might well be regarded as suspect. Furthermore, as we have argued, there are cases in which, while there is no outright inconsistency, there is such an overwhelming probability of falsehood that we must draw the line. Schotch and Jennings provide no help for this case.

VIII

It seems that we thus have good reasons to be skeptical of the logical rule of adjunction: from A and B to infer $A \& B$, except when it is applied to sentences at least one of which is a logical truth (and then, of course, we are in a position to employ a single empirical premise: $B \supset A \& B$). This is a syntactical constraint. It need not be paralyzing: in many cases when we are inclined to represent an argument as proceeding from a number of premises to a conclusion, we have perfectly adequate grounds for accepting the conjunction of the premises. Furthermore, we need make no adjustments to our classical semantics. It is not, for example, that the premises of the lottery paradox (ticket i will not win; the lottery is fair) do not *entail* its contradictory conclusion, but that the grounds we have for accepting each premise do not yield grounds for accepting the conjunction of the premises, and thus do not force us to accept the proposition that no ticket will win. More interestingly and more to the point in science, however low we set the parameter α of false rejection of the null hypothesis, we can be quite sure that in a large enough body of scientific results, the null hypothesis will have been falsely rejected at least once, though every individual rejection is deserving of our epistemological allegiance, and we have no way of singling out the false rejections.

For the same reasons, a deductively valid argument from premises we accept individually is, and should be, regarded as no more sound than the soundness of the argument from these individual premises to their conjunction. None of this undermines the classical philosophical view that, if the premises of an argument are known, and the argument is valid, then the conclusion must be true. If knowledge is taken to entail truth, then this view is perfectly correct. Its correctness, however, is not as interesting as it is often taken to be if we cannot identify premises as true. A knowledge *claim* is not the same as knowledge. What we have to work with in the real world are knowledge claims, reasonable beliefs. To have warrant for a number of knowledge claims is not at all to have warrant for the corresponding conjunctive knowledge claim. A deductive argument from a proposition reasonably believed transmits at least as much warrant to its conclusion as there is for its premise. A deductive argument from a *set* of propositions may provide no more warrant for its conclusion than there is for the *conjunction* of the premises. It is not unreasonable to suppose the warrant for a conjunction is no greater than the warrant for the least supported conjunct.

I have tried to argue that weak deductive closure, in which the entailments of any single acceptable premise are acceptable, and full deductive closure are very different things, and that the difference is important to the practical import of deductive arguments from premises. This applies to doxastic commitment as well as to argument. It is not at all clear that one should be obliged to believe the conjunction of items one believes, and the examples we have adduced suggest that one should not be so obliged. When it comes to what David Israel[22] has called *real logic*, the unsoundness of adjunction is exactly what divides real-world inference from the ideal inference of the classical logicians; and so it should.

HENRY E. KYBURG, JR.

University of Rochester

---

[22] "What's Wrong with Nonmonotonic Logic?" *Proceedings of the First Annual National Conference on Artificial Intelligence* (1980): 99-101.

FALLIBILISM, AMBIVALENCE, AND BELIEF

> The *intense* view of these manifold contradictions and imperfections in human reason has so wrought upon me, and heated my brain, that I am ready to reject all belief and reasoning, and can look upon no opinion even as more probable or likely than another. Where am I, or what? From what causes do I derive my existence, and to what condition shall I return? Whose favour shall I court, and whose anger must I dread? What beings surround me? and on whom have I any influence, or who have any influence on me? I am confounded with all these questions, and begin to fancy myself in the most deplorable condition imaginable, inviron'd with the deepest darkness, and utterly depriv'd of the use of every member and faculty.
>
> Most fortunately it happens, that since reason is incapable of dispelling these clouds, nature herself suffices to that purpose, and cures me of this philosophical melancholy and delirium, either by relaxing this bent of mind, or by some avocation, and lively impression of my senses, which obliterate all these chimeras. I dine, I play a game of back-gammon, I converse, and am merry with my friends; and when after three or four hour's amusement, I wou'd return to these speculations, they appear so cold, and strain'd, and ridiculous, that I cannot find in my heart to enter into them any farther.
>
> —David Hume, *A Treatise of Human Nature*

> I'm not so sure it's good to think back to my childhood memories, because I end up feeling happy and sad at the same time, and that gives me a weird "neutral" feeling.
>
> —Jack Handey, *Fuzzy Memories*

No concept plays a more central role in epistemology than the concept of *belief*. I offer here an account of what it is to believe a proposition. More precisely, what I am offering is an explication in Rudolf Carnap's[1] sense: a precise successor concept that captures significant features of our imprecise pretheoretic notion of belief. An explication is successful, Carnap tells us, if the explicatum (the successor concept) is similar to the explicandum (the pretheoretic concept), characterizable in a precise and simple way, and fruitful in its application to theoretical problems. It is not required that the explicatum cleave precisely to every nuance of usage associated with the explicandum in everyday discourse; given the vagueness and fuzziness that infect most of the concepts expressible in ordinary language, it is often more useful to permit significant divergence from lay usage in order to achieve greater precision and simplicity. This is especially likely to be

---

[1] *Logical Foundations of Probability* (Chicago: University Press, 1962, 2nd ed.).

0022-362X/97/9403/126–55

the case when the explicandum is as broad and multifaceted as the ordinary concept of belief. Given these considerations, I want to make it clear from the outset that I am not attempting anything like an analysis of belief in the widest sense of the term; my discussion will be limited to a single particular aspect of the everyday concept. I am pursuing this strategy not because I regard other aspects of the concept as unimportant, but because I think our overall understanding of belief will be best served by carving the concept at its joints, replacing it with several related but distinct explicata designed to serve the various realms of theoretical discourse in which we talk about belief.

In order to get at the aspect of belief about which I want to talk, I shall begin by describing an important aspect that I intend to exclude. One of the most important ways in which we use the notion of belief in ordinary discourse, and certainly the way that has attracted the most philosophical attention in recent years, is as a component of our commonsensical psychological theories. We attribute beliefs to other people (and, significantly, to animals as well) in order to explain their past behavior and predict their future behavior. In other words, we often talk about beliefs as if they are psychological states that play a distinctive and significant role in the etiology of human and animal activities. As such, they are similar to the theoretical entities posited by other branches of empirical science in order to systematize and explain complex patterns of observable phenomena. If we approach the explication of belief from this perspective, we shall be inclined to clarify our understanding of belief in the same way we clarify our understanding of other posits of empirical science, namely, by clarifying our understanding of the theoretical framework in which the concept is embedded. On this approach, we explicate belief by describing the psychological laws that govern the regular patterns of interaction between beliefs and other psychological states that are responsible for the production of behavior. Belief is defined in terms of the functional role it plays according to our best psychological theories. Although the claim has been disputed, it is not implausible to think that such an analysis will make detailed reference to the cognitive architecture that underlies human and animal psychological processes. We might, for example, end up characterizing the state of believing that p in terms of the tokening of a sentence that states that p (in "mentalese," as it were) within a cognitive system designed to use such sentence tokens as data in computations that produce appropriate directives for behavior.[1]

---

[1] See Jerry Fodor, *The Language of Thought* (New York: Crowell, 1975) for the classic defense of this approach, and Daniel C. Dennett, *Brainstorms* (Cambridge: MIT, 1978) for a powerful attack on the claim that theories of belief must be cashed out in terms of underlying cognitive architecture.

Without disputing the importance of this aspect of our concept of belief, I want to stress that such analyses do not exhaust the concept. The point is most readily seen if we look at first-person ascriptions of belief. As Bas C. van Fraassen[3] notes, sentences of the form 'I believe that *p*' are often not comfortably read (in the first instance) as episodes of autobiographical description in which one attributes a psychological disposition to oneself. Such statements are more plausibly regarded as speech acts similar to the making of a promise or the declaration of an intention: by declaring that one believes that *p*, one undertakes certain normative commitments, commitments that govern how one ought to engage in future deliberation and inquiry if one intends to pursue these activities in good faith. Likewise, we often use third-person belief ascriptions not to offer a causal explanation of another's behavior but to indicate that the party in question is explicitly or implicitly committed to the truth of a proposition in the appropriate way. Philosophical epistemology, inasmuch as it is concerned with the appropriateness of various practices of deliberation and inquiry, is chiefly concerned with beliefs in this sense. Of course, the psychological and epistemological aspects of belief are not wholly independent; one can often infer much about the psychological state of an individual from information about her epistemic commitments, and vice versa. But the two notions are clearly separable; note that beliefs in the epistemic sense can be attributed to entities, such as corporations and societies, that do not even have psychological properties in the ordinary sense.[4]

It is worth noting the work of two other authors who have drawn distinctions roughly parallel to the one I have sketched above. L. Jonathan Cohen[5] defines *acceptance* of a proposition as voluntary adoption of a policy to use that proposition as a premise in one's reasoning. Similarly, Daniel C. Dennett defines *opinions* as the product of decisions to assent to sentences, roughly akin to gambles that those sentences are true (*op. cit.*). Both authors contrast these concepts with the notion of belief, where the latter is construed as something like an involuntary psychological disposition to feel that a proposition is true and to behave accordingly. Neither acceptance nor opinion corresponds precisely to the epistemic sense of belief I am carving out. Cohen's concept is broader than mine; it includes, for example, the sort of commitment that a lawyer undertakes with respect to the innocence of his client or that a scientist undertakes with respect to an admittedly controversial but favored pet theory, commitments that I am reluctant to categorize as instances of genuine belief. For Dennett, opinion is necessarily a linguistic phenomenon, whereas I want to allow that an agent might be committed to more beliefs than are expressible in any given language. Nevertheless, there are clearly important areas of overlap among these three notions, and much of what those authors have to say strikes me as true of, and relevant to, the analysis of the aspect of belief with which I am concerned.

---

[3] "Belief and the Will," this JOURNAL, LXXXI, 5 (May 1984): 235-56.

[4] Margaret Gilbert, "Modelling Collective Belief," *Synthese*, LXXIII (1987): 185-204, provides an illuminating analysis of this sort of group belief.

[5] *An Essay on Belief and Acceptance* (New York: Oxford, 1992).

Dividing up the concept of belief in this way allows us to bypass some of the stumbling blocks that have plagued much philosophical discussion of the propositional attitudes. Chief among these is the specter of eliminativism. Folk psychology is an empirical theory; moreover, it is a theory of rather dubious intellectual heritage. In this respect, it resembles many of our other folk theories, such as the common-sense Aristotelian physics which, we all have to be taught in school, is wildly false in many aspects. So it is an open possibility that the progress of scientific psychology will result in the development of theoretical frameworks for the explanation and prediction of behavior which leave no room for entities like the beliefs we posit in our everyday psychological practice. It may well turn out, in other words, that beliefs (in the folk-psychological sense) just do not exist.[6] Given this possibility, we might well fear that any philosophical analysis of belief will be as empty as an academic discussion of black magic or a treatise on the nature of phlogiston. As long as our attention is restricted to belief in its epistemic sense, however, such worries need not bother us. That individuals are capable of undertaking normative commitments to engage in deliberation and inquiry in certain ways is not a fact that is in imminent danger of being refuted by new developments in cognitive psychology. The structure of our epistemic practices is a matter of personal and cultural decision, not of our innate cognitive architecture; and as long as we find belief talk useful in the conduct of inquiry, we are entitled to preserve it as a central component of that conduct without fear of violating the dictates of good scientific practice.

One author who has been persistently mistaken on this point is Paul M. Churchland.[7] He is one of the most vocal defenders of eliminativism; he insists that the progress of neuroscience will soon force the displacement of our body of common-sense psychological concepts. Moreover, he has argued that this revision will entail the necessity of a parallel revision in the concepts that govern epistemology and the philosophy of science; once it becomes clear that there are not any beliefs in our heads, we shall be forced to reframe our methods for evaluating deliberation and inquiry within the new theoretical apparatus of neuroscience. He writes:

---

[6] John D. Greenwood, ed., *The Future of Folk Psychology* (New York: Cambridge, 1991) provides a useful survey of the available philosophical positions on this issue.

[7] *A Neurocomputational Perspective* (Cambridge: MIT, 1989).

> Autonomy [of epistemology] must be denied because normative issues are never independent of factual matters.... We speak of *justification*, but we think of it as a feature of *belief*, and whether or not there are any beliefs and what properties they have is a robustly factual matter. We speak of *rationality*, but we think of it as a feature of *thinkers*, and it is a substantive factual matter what thinkers are and what cognitive kinematics they harbor. Normative concepts and normative convictions are thus always hostage to some background factual presuppositions, and these can always prove to be superficial, confused, or just plain wrong. If they are, then we may have to rethink whatever normative framework has been erected upon them. The lesson of the previous pages is that the time for this has already come (*ibid.*, p. 196).

But I do not see any reason to think that the concepts Churchland mentions, in the sense in which those concepts are employed in epistemology, embody any controversial factual assumptions at all. Again, it is simply not in the cards that neuroscience might one day convince us that there is something ill founded or incoherent about the practice of committing oneself to the truth of a proposition, any more than it is a possibility that developments in quantum physics will force us to conclude that no one has ever actually made a promise. This is not to say that scientific psychology can make no contribution to epistemology. We are not bound to our present conceptual system; we are free to revise it in any way that strikes us as useful, and neuroscience is as good a source as any for inspiring new developments. But the criterion for revision is usefulness with respect to our purposes as we conceive them, not faithfulness to some scientistic ideal. Until Churchland has some genuine suggestions about how to use the concepts of neuroscience to improve our practices of evaluating deliberation and inquiry, and not just a sketch of a utopian program that insists that such improvements are forthcoming, we need not worry that our discussions of belief are in imminent danger of being rendered obsolete.

Carving up belief in this way also affects the position we ought to take on the acceptability of various putative normative principles that govern the adoption and revision of one's beliefs. For present purposes, the most important such maxim is the principle of deductive cogency. According to this rule, one ought to believe all of the deductive consequences of any set of propositions all of which one believes; furthermore, one ought not believe any contradictory set of propositions. How plausible is this normative claim? If we are thinking of beliefs as something like sentences stored in the brain, not plausible at all. Gilbert Harman[8] is surely right in maintaining that a well-designed cognitive system will embody a policy of clutter avoidance: such a system will not waste valuable time, storage space, and resources generating complicated tautological consequences of sentences already stored or testing sets of stored sentences for mutual compatibility. If, however, we are thinking of beliefs as epistemic commitments, the situation is reversed. It is no difficult feat to undertake even an infinite number of commitments all in one stroke; all it takes is a simple act of will. There is, of course, no normative requirement that one keep an explicit record of all one's commitments, nor that one be able to enumerate those commitments upon demand, nor that one be able in all cases to recognize spontaneously that one has violated those commitments. All that is required is that one not knowingly regard a violation of one's commitments with equanimity: if one is made aware that one has violated one's own commitments, one must change either one's practice or one's commitments to remain in good faith. Given this perspective, it seems entirely reasonable to postulate that an agent must believe any proposition that is implied by that agent's other beliefs. Thus, in what follows, I shall typically assume that an agent's beliefs can be represented by a single (nonempty) proposition. This proposition, the agent's *corpus*, is the conjunction of all of the propositions believed by the agent; the agent's beliefs are all and only the deductive consequences of the corpus.

One more remark with respect to clarification of the explicandum, and we shall be ready to proceed. We sometimes think of beliefs as relations to sentences, sometimes as relations to propositions, and sometimes as relations to both at once. In what follows, I shall in general regard beliefs as purely propositional attitudes. The intuitive picture I have in mind is something like the following: the epistemic agent is confronted with a space of logical possibilities. (For present purposes, it does not matter to me whether these possibilities are possible worlds, possible individuals,[9] or some other sort of entity.) The agent's apprehension of these possibilities is, as it were, direct; it is not mediated by some form of linguistic or pictorial representation. (An idealization of our own epistemic situation, to be sure; but not a dangerous one, and close enough to the truth to suit our purposes for a wide variety of cases.) The epistemic task that the agent faces is the division of the possibility space into two subspaces: one consisting of possibilities included in the corpus, the other of possibilities excluded by the corpus. If we define propositions to be subsets of the set of possibilities, then the agent's beliefs will be just those regions of the possibility space which include the corpus as a (perhaps improper) subregion. Note that this way of thinking of belief makes it problematic to talk about some familiar cases of belief. Logical beliefs—beliefs about which sentences in a language express the necessary proposition (the entire possibility space) and which express the contradictory proposition (the empty set)—are not comfortably representable in this system; nor are modal beliefs—beliefs about the structure and content of the possibility space. (I take it that mathematical beliefs by and large fall into one of these categories.) I am uncertain what modifications would be required to allow my analysis to accommodate these types of belief. Nevertheless, my explication is still valuable as an account of that category of belief most central to our primary epistemic activities, namely, empirical beliefs about the way the world actually is.[10]

---

[8] *Change in View* (Cambridge: MIT, 1986).

[9] See David Lewis, "Attitudes *De Dicto* and *De Se*," *Philosophical Review*, LXXXVIII (October 1979): 513-43.

I

I hope I have said enough at this point to give you a reasonable sense of what concept I am attempting to explicate. Let us turn, then, to the explicatum. In terms of what other concepts is it appropriate to characterize the nature of epistemic belief? My proposal is that the central notion we need is *serious possibility*, a concept that plays a major role in the philosophy of Isaac Levi.[11] I shall be discussing Levi's views at some length, primarily because his position serves as a useful foil to my own. He, like me, thinks that the concept of belief can be illuminated through explication in terms of serious possibility; his analysis, however, differs from mine in significant and instructive respects.

The concept of serious possibility, while presumably not wholly unfamiliar, is not as familiar and central as the concept of belief. I take it, though, that this poses no difficulty for the explication of the latter in terms of the former; the philosophical value of analyses that proceed from more to less basic concepts is, one hopes, sufficiently well established so as not to require defense here. In introducing the concept, I can do no better than to quote Levi directly:

---

[10] For further elaboration of this intuitive picture and discussion of related issues, see Robert Stalnaker, *Inquiry* (Cambridge: MIT, 1984).

[11] See *The Enterprise of Knowledge* (Cambridge: MIT, 1980), and *The Fixation of Belief and Its Undoing* (New York: Cambridge, 1991).

> When witnessing the toss of a coin, *X* will normally envisage as possibly true the hypotheses that the coin will land heads up and that it will land tails up. He may also envisage other possibilities—e.g., its landing on its edge. However, if he takes for granted even the crudest folklore of modern physics, he will rule out as impossible the coin's moving upward to outer space in the direction of Alpha Centauri. He will also rule out the hypothesis that the Earth will explode.
>
> Of course, it is logically possible that the coin will fly out towards Alpha Centauri upon tossing and even that the Earth will explode under these conditions. But this means only that if *X* were in that (different) cognitive state wherein his standard for serious possibility consists exclusively of logical truths, these hypotheses would be serious possibilities according to *X*....
>
> Logically possible though these hypotheses may be, *X* both would and should ignore them because they are not seriously possible from his point of view....
>
> The distinction between logical possibilities which are serious and those which are not is important in scientific inquiry as well as practical deliberation. When devising hypotheses about the constitution of quasars, no one considers the logical possibility that they are conglomerations of drosophilia flies to be a serious possibility. In designing experiments, one takes into account only those hypotheses as to the outcome which are serious possibilities.[12]

In other words, to regard a proposition as seriously possible is to adopt a policy of allowing that proposition to play a significant role in one's practical and theoretical deliberations, that is, to give the logical possibility that the proposition is true some nonnegligible weight in determining the outcome of one's decisions. By contrast, a proposition that is ruled out as not seriously possible has no impact on the conduct of inquiry; an agent who takes a proposition to be seriously impossible will follow just the same course of reasoning and decision making as an (otherwise identical) agent for whom that proposition does not lie within the space of logical possibilities at all.

Levi assumes that a rational inquirer will, at any given time, be committed to a single standard for serious possibility, that is, a division of the logical possibility space into two regions, one containing those logical possibilities which are taken seriously and the other containing those possibilities which are ruled out. In other words, the standard for serious possibility can be represented by a single proposition, namely, the proposition that contains all and only those possibilities which are serious. Propositions that are entailed by the standard (that is, which contain it as a subset) are seriously necessary; propositions that are compatible with the standard (that is, which overlap it) are seriously possible; and propositions that are disjoint from the standard are seriously impossible. It is easy to see that the seriously necessary propositions are subject to the requirement of deductive cogency. If this model of serious possibility strikes one as similar to the model of belief in terms of the selection of a corpus proposition, this is no accident. For on Levi's account, an agent's corpus and standard for serious possibility will be one and the same proposition. This is the core of Levi's analysis of belief: to believe a proposition is to regard it as seriously necessary, that is, to rule out alternatives to that proposition as not seriously possible. For Levi, judgments of believability and judgments of serious possibility coincide.[13]

---

[12] *The Enterprise of Knowledge*, pp. 3-5.

On first inspection, Levi's proposal may seem to have much to recommend it. After all, there is surely something right about the idea that to believe a proposition is to cease to regard alternatives to that proposition as serious candidates for the truth. Further reflection, however, reveals some disturbing consequences of Levi's development of this idea. Do we really want to say that a rational inquirer ought never to allow the course of her deliberation to be swayed by consideration of the possibility that propositions she believes to be false might nevertheless be true? To give an affirmative answer to this question is to adopt a position of *infallibilism*, that is, to maintain that one ought to regard all of one's beliefs as certainly and infallibly correct. Levi clearly recognizes this consequence of his position; indeed, he regards it as a central feature of his analysis. He writes:

> An immediate consequence of the thesis that *X*'s corpus at [time] *t* serves as his standard for serious possibility at *t* is that, according to *X* at *t*, no item in his corpus at *t* is possibly false in the sense of serious possibility. If $h \in K_{X,t}$ [*X*'s corpus at *t*], then *h* is infallibly true according to *X* at *t* in a straightforward and important sense. Thus if *Y* should disagree with *X* at *t*, from *X*'s point of view, *Y* is certainly in error. It would be in-

---

[13] In *The Enterprise of Knowledge*, the concept he takes himself to be explicating is *knowledge*, not belief. But his remarks about this concept make it clear that he is considering knowledge ascriptions from a first-person point of view, a perspective from which knowledge collapses into belief. This reading of Levi is borne out by the fact that in *The Fixation of Belief*, he switches terminology from knowledge to belief without altering the analysis.

> consistent for *X* to concede to *Y* that he (*X*) might be mistaken if by this *X* acknowledges the falsity of *h* as a serious possibility.
>
> Of course, *X* may consistently acknowledge that items he accepted in his corpus at previous times or will assume in the future are possibly false. He is not committed to the view that whatever he has endorsed in the past or will accept as evidence in the future is infallibly true. But at *t*, *X* is committed to the view that whatever he assumes as part of his corpus at t is infallibly true.[14]

But I take this consequence to be deeply troubling. Most of us, I think, are inclined to adopt a position of *fallibilism* with respect to our own beliefs; we are willing to concede that there is some non-negligible likelihood that some of our opinions are mistaken. At minimum, I take it that most of us will allow that fallibilism is at least a coherent position which a rational agent may adopt. The argument for fallibilism is the simple argument from humility and mortal imperfection: I take it for granted that virtually everyone I know has some false beliefs, and I regard it as almost certain that the belief systems of most of the people alive in previous eras were wildly false with respect to nearly every question of intellectual significance. Why should I exempt myself from the obvious inductive conclusion? Likewise, as Levi grants, we generally take seriously the possibility that our past or future belief corpuses contained errors. But there is nothing epistemically privileged about my present position, so there seems to be no reason why I should not take just as seriously the possibility that my present corpus is erroneous in certain respects.

We can confirm that the rationality of fallibilism is supported by our intuitions by considering the judgments we are inclined to make with respect to certain practical decision problems. This point is usually made via reference to a wager at incredibly high stakes. Consider some proposition I believe to be true—for example, the proposition that Millard Fillmore was born on January 7. Suppose that someone offers me the following bet: if this proposition is true, I shall receive a prize of almost but not entirely negligible worth, such as a two-for-one coupon from Burger King; if, on the other hand, the proposition is false, I shall incur a penalty of massive disutility, such as an excrutiatingly painful and permanently disfiguring regimen of physical torture. On Levi's infallibilist account, I am rationally compelled to accept this wager. This is because, from his perspective, I must regard the falsehood of the proposition in question as not seriously possible; hence, contemplation of the possibility that I shall incur the penalty, however undesirable that consequence seems to me, can play no significant role in my decision. The only factor that rationally matters is the small net profit I regard myself as certain to gain. But I feel fairly confident that most of us do not share this judgment. Surely, it is rationally permissible to reject the wager. This conclusion, however, can only be justified if we allow that I give serious consideration to the possibility that I shall lose the wager if I take it.

---

[14] *The Enterprise of Knowledge*, p. 13.

Some philosophers have questioned the cogency of this form of argument, suggesting that the introduction of utilities and disutilities of extraordinary magnitude can have a distorting effect on our intuitions. Therefore, I think it is worth reinforcing the point by considering a class of decision problems that have been less widely discussed in the philosophical literature. These cases, which revolve around the phenomenon of hedging one's bets, also demonstrate the rationality of fallibilism, but do so without introducing extreme utilities. An example: suppose that I am a contestant on a *Jeopardy*-style television quiz show. I have amassed a bankroll of eight thousand dollars so far in the competition, and I now have an opportunity to wager some or all of this sum on my ability to answer the question: "On what date was Millard Fillmore born?" Recall that I believe that Millard Fillmore was born on January 7. Levi's system, then, entails that there is exactly one rational course of action available to me, namely, wagering the entire eight thousand dollars. The reasoning is the same as above: since it is not seriously possible that my answer to the question will be incorrect, considerations of potential loss ought to play no role in my decision, and I am compelled to choose the option that maximizes my potential profit. Once again, however, this judgment appears to violate our intuitions. Most of us would grant that it is rationally permissible for me to wager less than my full bankroll, reserving some of my money as insurance against the possibility that my answer will turn out to be wrong. Again, this response only makes sense on the assumption that I take the falsehood of my belief to be a possibility worthy of serious consideration. The result holds even though nothing of particularly momentous importance hangs on the outcome of my decision.

Infallibilism is also problematic with respect to purely theoretical deliberation as well as practical decision making. To take just one example, consider the long history of philosophical debates over skepticism in its various manifestations. To be sure, a number of distinct doctrines, questions, and confusions have been subsumed under the label of "skepticism." It seems unproblematic, however, to make the claim that at least one of the issues that have played a significant role in the debate over skepticism is the question of how seriously to take the possibility that most or all of our beliefs in a given domain are radically false. Different thinkers have staked out different positions on this question. To enter seriously into the debate over skepticism at all, however, it seems necessary to presuppose that it is at least conceivable that one could take skeptical hypotheses seriously. Skepticism can only be a serious philosophical issue if it is not the case that the nature of belief rules out as conceptually incoherent the act of seriously doubting that one's own beliefs are true. To adopt Levi's infallibilist analysis of belief, then, is effectively to dissolve the entire question of skepticism. But this move cannot be acceptable to any philosopher who regards the issue of the rational force of skeptical hypotheses to be a serious and open question. Since I rank among these philosophers, I cannot find Levi's account acceptable.

But the most serious challenge facing infallibilism is the worry that it collapses into *incorrigibilism*. Most of us, Levi included, take inquiry to be a rational activity. That is, we take it that we may rationally engage in intellectual pursuits that are designed to produce revisions in our belief systems; we actively seek out new evidence and engage in arguments with opponents in the knowledge that such activities are likely to cause us to discard some of our present beliefs and replace them with new ones. We adopt a position of *corrigibilism* with respect to our present beliefs. The question that faces us is: Can an infallibilist such as Levi coherently endorse corrigibilism? On the face of it, the answer seems to be negative. If I genuinely rule out alternatives to my present beliefs as not seriously possible, then the only appropriate methodology for me to adopt seems to be what C. S. Peirce called the *method of tenacity*. I ought actively to avoid any exposure to new experimental results, new arguments, or anything else that might lead me to revise my beliefs. After all, my present beliefs are certainly true, while the beliefs of my opponents have no possibility of being true; so any process that causes me to abandon my beliefs and adopt theirs must rely on some form of trickery or deception. Unless I regard my own beliefs as possibly false, inquiry seems to be pointless at best, dangerous at worst. Incorrigibilism is the only tenable response.

Levi takes the incorrigibilist challenge very seriously; he devotes a good deal of effort to its resolution. But the solution he arrives at is, I think, ultimately indefensible. Levi grants that it is indeed irrational for an agent to decide to shift directly from belief in a proposition p to belief in its negation $\neg p$; from the perspective of the agent prior to the shift, such a move represents deliberate abandonment of certain correctness in favor of certain error, and is thus to be avoided at all costs. He argues, however, that such a shift may be justified if it is carried out indirectly as the net product of two smaller changes. First, the agent suspends judgment between p and $\neg p$. Since this involves only the abandonment of a prior belief, not the adoption of a new belief, such a move will not result in the importation of certain error into the agent's corpus and is therefore not automatically disqualified. After this shift, the agent will judge serious possibility from a new epistemic perspective such that $\neg p$ is now regarded as a serious possibility. Thus, the agent will no longer regard the addition of $\neg p$ to the corpus as problematic, since she is no longer infallibly certain that it is false. What cannot rationally be accomplished in one fell swoop may be undertaken incrementally.

One might question, however, the reasonability of the first of the two shorter steps, in which the agent contracts her corpus to exclude p. Granted, such a move does not import any falsehoods into the corpus, but it does result in the loss of a certainly true piece of knowledge. Why should the agent willingly sacrifice information in this way? Levi suggests that the loss may be compensated by the gain of other epistemic virtues:

> Suppose the initial corpus contains some theory T_1. A second theory T_2 contradicts T_1. From X's initial point of view, T_2 is certainly false. Yet it may be superior in all other respects to T_1 as a means for furnishing systematic explanations in some domain. X could recognize the superior explanatory virtues of statistical mechanics even though he is certain that it is false and that classical thermodynamics is true.
>
> In such cases, X might be prepared to suffer a loss of information due to the removal of T_1 from his corpus in order to be in a position to take the truth of T_2 to be seriously possible. In that event, X exchanges a given amount of information for the opportunity to give an informationally attractive hypothesis a hearing.
>
> To contract in this case is not to reject T_1 as false and to accept T_2 as true but to shift to a position where judgment is suspended between the rival hypotheses so that investigation can be undertaken to decide whether T_1 should be reinstated via inferential expansion or T_2 should take T_1's place (*ibid.*, p. 60).

But the suggestion is inadequate as it stands. Ordinarily, we do not take the superior explanatory virtues of a theory to provide sufficient reason by themselves for giving that theory a fair hearing. Consider, for example, an ancient cosmology that subsumes a wide variety of natural phenomena under a handful of simple explanatory schemas, most of which refer to the behavioral dispositions of a small group of anthropomorphic beings of vast power whose psychological traits are antecedently specified. There is no question that such theories can have significant explanatory power. Someone equipped with a good working knowledge of contemporary science could probably devise a new pantheon of deities whose activities could serve as the basis of an even more comprehensive and accurate explanatory system. Similar remarks could be made about various folk legends, science fiction stories, astrological and numerological systems, and a vast array of other theories of dubious repute. Why do we not waste our time giving theories such as these a proper hearing? From my perspective, a simple answer can be given: I am unwilling to suspend judgment with respect to the truth of, say, Babylonian mythology, despite its explanatory power, because I do not regard the truth of Babylonian mythology as a serious possibility. There are other scientific hypotheses, such as the existence of cold fusion, which I believe to be false but which I nevertheless regard as seriously possible; inquiry concerning these hypotheses strikes me as worthwhile in a way that research seeking to establish the truth of Babylonian mythology does not. But Levi, of course, cannot draw these distinctions. For him, all alternatives to one's present beliefs are seriously impossible; hence there is no ranking them with respect to comparative plausibility. Absent such plausibility judgments, Levi owes us an account of how to distinguish between those certainly false hypotheses which deserve serious consideration and those certainly false hypotheses which are just plain silly.

An even more serious problem faces Levi's two-stage account, however. Consider the position of an agent trying to decide whether or not to drop her belief in p in order to give $\neg p$ a fair hearing. Such an agent presumably ought to take into account all of the foreseeable consequences of the available options. But suppose it is clear to the agent that, if she comes to regard $\neg p$ as seriously possible, she will take its other virtues (such as superior explanatory power) to provide sufficient reason to add $\neg p$ to her corpus. In other words, she knows from the outset that taking the supposedly innocuous first step of the sequence will yield the same net result as shifting directly from belief in p to belief in $\neg p$. Under such circumstances, it seems clear that the agent ought to eschew the proposed belief contraction for the same reason that she rejects the direct shift: such a decision will ultimately result in the abandonment of certain truth for certain error. Dividing the process into two steps becomes a mere formal trick which in no way alters the incorrigibilist conclusion.

Levi's response to this challenge is somewhat curious. He stakes out a position that he calls *secular realism*, a position that denies crucial assumptions about the goals of inquiry that motivate the above challenge:

> According to secular realism, the inquiring agent is committed to maximal certainty about the truth of the assumptions in his current corpus. He is concerned, moreover, to avoid error in the next alteration he makes in that corpus, where error is judged on the assumption that everything in the current corpus is true. Of course, the agent can recognize that in making changes in the current corpus, he might and probably will end up several stages down the chain of belief revision, becoming certain of claims he currently is certain are false. But as a secular realist, he is not concerned with the possibility that his beliefs several stages down the chain will be false as he currently judges truth and falsity. He is not concerned with progress toward the true, complete story, but only with avoiding error at the next step.[13]

Levi argues for secular realism largely by contrasting it with an alternative position he calls *messianic realism*, a thesis that he identifies with the epistemological outlooks of Peirce and Karl Popper. According to messianic realism, "truth is a regulative ideal directing us toward the End of Days, when inquirers will have attained the true, complete story" (*ibid.*, p. 164). Messianic realists are thus concerned with the avoidance of error not just at the next stage of inquiry, but at all subsequent stages as well, or perhaps just at some ultimate stage when science will have converged on the ideal "theory of everything." Levi heaps derision on messianic realism, and he suggests that messianic realists can give no account of the role of truth seeking and error avoidance in proximate and short-term inquiry (although his argument for this conclusion is obscure to me).

At any rate, the contrast Levi offers us is a false one; we need not saddle ourselves with the messianic position in order to recognize that there is something wrong with secular realism. We might, for example, stake out an intermediate position—*presbyterian realism* seems an appropriate name—which makes no claims at all about the end of days or the ultimate goals of science, but which holds that avoidance of error is a desideratum in inquiry not just at the next stage, but also at those stages in the immediate future whose outcomes are foreseeable with a reasonable degree of accuracy. Presbyterian realism supports the assumption that underlies the incorrigibilist challenge to infallibilism, namely, that there is no real epistemic distinction between a one-step and a two-step importation of error into one's corpus. In light of the availability of this position, what we require from Levi is not more invective against messianic science worship, but a general argument that the peculiar form of intellectual myopia he recommends really is a rational approach to inquiry. And such an argument does not seem to be forthcoming. Certainly, secular realism cannot be derived from a general prohibition against taking future belief revisions into consideration in the course of one's deliberations. Suppose, for example, that I am offered a potent drug, the ingestion of which will cause me a good deal of immediate pleasure, but which will lead some weeks from now to the development of a susceptibility to suggestions that I can safely leap from tall buildings, stop trains with my bare hands, and so forth. Without a doubt, it is rational for me to refuse this drug precisely because it will eventually cause me to adopt beliefs I now regard as dangerously false. So in the general case, rational decision making requires me to consider the effect of my decisions on belief revisions other than the most proximate one. Why should this requirement disappear when the decision in question is epistemic rather than practical? Levi's only real argument is that secular realism is required in order to reconcile infallibilism with corrigibilism. To which I can only reply: so much the worse for infallibilism.

---

[13] *The Fixation of Belief and Its Undoing*, p. 163.

II

Levi's account of the relationship between belief and judgments of serious possibility is clearly untenable. Nevertheless, as I have already indicated, the basic intuition that what one believes is ultimately a matter of what one regards as seriously possible deserves serious attention. It is time for me to put my cards on the table and spell out my own proposal about the real relationship between one's corpus and one's possibility standard.

As a first step toward the proposed explication, I want to focus our attention on a particular autobiographical example. As I have mentioned before, among the propositions I believe is the proposition that Millard Fillmore was born on January 7. What is my attitude toward this proposition with respect to my judgments of serious possibility? As I reflect on this question, I find that my opinion on the matter exhibits a genuine instability. On Mondays, Wednesdays, and Fridays, as it were, I am inclined to regard the proposition as seriously necessary. January 7 *is* Millard Fillmore's birthday, after all; there is no real controversy on the matter, and no real reason to speculate about the possibility that other dates might be more correct. The issue is settled, and deliberating further on it is just pointless. On Tuesdays, Thursdays, and Saturdays, on the other hand, I am inclined to allow that there are genuine grounds for doubt as to the correctness of my belief. Consider: my only evidence for taking January 7 to be Fillmore's birthday is an article I read in the *World Book Encyclopedia* nearly fifteen years ago. But my memory is not perfect; neither is the editorial staff at the *World Book*. I have had plenty of beliefs in the past which were much better supported than this one but which ultimately turned out to be false. So when I am in this frame of mind, I am inclined to regard alternatives to January 7 as serious candidates for Fillmore's actual birthday.

What are we to make of this instability? From Levi's perspective, the appropriate thing to say is that, contrary to my initial claim, I do not really have an opinion on the matter of Fillmore's birthday. Rather, I have two opinions, and a tendency to vacillate between them at distressingly frequent intervals. I am perpetually adding the belief that Fillmore was born on January 7 to my corpus, only to turn around and reject this belief once again. But this description just does not seem to me to be true to the facts or to our ordinary way of talking about these matters. My vacillation on this question does not strike me as of a piece with the genuine changes of mind I periodically undergo, such as when I acquire new information or accept the consequences of a novel argument. Rather, it seems to me that both of my stances on this issue are partial reflections of a single deeper state of opinion, a state which has remained fairly stable over the last fifteen years, my vacillations notwithstanding. Psychological factors, such as the immediate salience of certain facts or arguments, may make one facet of that opinion strike me as momentarily compelling; but if I step back from these distorting factors and enter a more self-reflective state, I find myself partially identifying with both judgments simultaneously.

The phenomenon of epistemic ambivalence that I am describing has been identified and discussed by a number of philosophers of note. The epigraph from Hume with which I began this discussion represents a good example. Hume's position on philosophical skepticism seems to be that we are all naturally inclined to regard skeptical theses with the same ambivalence I have described myself as adopting toward the Fillmore question. Within the confines of the study, confronted with the full force of skeptical arguments, we are driven to adopt extremely broad standards for serious possibility, feeling ourselves at a loss to rule out anything whatsoever with any certainty. As soon as we change our immediate surroundings and thus our psychological set, however, we find it extremely difficult to take seriously the possibilities that gripped us before. The change is not driven by new evidence or superior reasoning; Hume does not think that skepticism can be dispelled by these means. It is driven instead by a mere shift in context, one that makes certain features of our judgments more immediately salient than others. Likewise, Peirce in his defense of fallibilism seems committed to something like this picture of ambivalence. On Peirce's extreme version of fallibilism, the scientific spirit requires that we never rule out any option at all as a serious possibility. To this extent, theoretical science is the enemy of practical deliberation, which requires that we rule out many logical possibilities in order to motivate immediate decisions to pursue courses of action. Yet Peirce allows that one and the same agent may inhabit the worlds of science and of practical affairs simultaneously, and presumably may do so without lapsing into bad faith with respect to one's epistemic commitments. Thus Peirce, like Hume, seems implicitly committed to the view that we may simultaneously hold different standards for serious possibility, some of which are more salient in purely theoretical reasoning and some of which present themselves most prominently in practical contexts.

Epistemic ambivalence also shares important features with emotional ambivalence, a phenomenon that has been given an insightful treatment by Patricia S. Greenspan.[18] She writes,

> [W]e may be said to have or exhibit a particular emotion (and indeed, I might add, to exhibit it consciously) over a span of time which includes, but is not limited to, the times (supposing there are some) when I am actually experiencing it. Thus, if I waver, over time, between happy and unhappy feelings about my rival's victory (I momentarily feel bad when I first hear the news on the phone, say; but then I immediately consider my friend's good fortune and momentarily feel good), we would reasonably conclude that I have "mixed feelings" throughout the overall time span involved, and not that I am continually changing my mind (*ibid.*, p. 229).

Exactly similarly, I want to suggest that I am at least partly committed to the view that Fillmore's birthday's being on January 7 is seriously necessary even during the periods of time when I do not find this view immediately compelling (and likewise for the view that other dates are seriously possible). It is this partial commitment which constitutes my belief that Fillmore was born on January 7. Just as my stable emotional state transcends the vagaries of my momentary experience, belief is a state of opinion that transcends the vagaries of my momentary psychological set. I do not believe a given proposition only during those moments in which I hold it with firm and absolute conviction, eschewing as patently false all alternatives; I can continue to believe a proposition even as I cycle through moments of certainty and doubt, of trust and skepticism. Consider the game show example once again: if I decide not to bet my entire bankroll on my ability to state Millard Fillmore's birthday correctly, this does not mean that I have suddenly ceased to believe that Fillmore was born on January 7. It merely means that the salient features of my immediate circumstances make me find the skeptical facets of my underlying opinion more compelling than the dogmatic facets.

---

[18] "A Case of Mixed Feelings: Ambivalence and the Logic of Emotion," in Amélie Oksenberg Rorty, ed., *Explaining Emotions* (Berkeley: California UP, 1980).

What all this suggests is that my state of opinion with respect to judgment of serious possibility is not adequately represented by a single standard that divides up the possibilities into just two categories; my judgments exhibit a more finely detailed internal structure than this model allows. We do better by supposing that I am committed to a set of propositions, each representing a standard of serious possibility with which I am willing to partially identify. Picking out this set, that is, identifying those judgments to which I am actually committed, requires a certain degree of mental discipline, since in general it requires me to step outside my momentary frame of mind, ignoring the contingencies of my psychological set which may be driving me to identify with a restricted subset of the available standards, in order to recognize the broader range of opinions that I am willing to call my own. It is no part of my account to claim that we are always, or even often, in a position to carry out this sort of self-transcendent reflection. Our relationship to these standards is the same as our relationship to all of our normative commitments: inasmuch as our limited capacities allow us to perceive clearly what obligations we are committed to, good faith requires us to carry out these obligations as best as we can, as judged by our own lights.

In connection with this last point, it is imperative that I clear up a possible confusion that my discussion up to this point has (to some extent deliberately) fostered. In order to identify clear examples of epistemic ambivalence, I have focused on cases in which two different states of opinion were each made psychologically salient in a given set of circumstances. But I do not wish to claim that such psychological immediacy is either a necessary or a sufficient condition for partially identifying with a given state of opinion. There may, for example, be standards for serious possibility that I recognize upon reflection as acceptable to me, even though there is no conceivable context I could find myself in within which those standards would strike me as more compelling than other standards with which I also identify. Likewise, I may make immediate judgments of serious possibility which momentarily grip me, but which from a position of self-reflection I decide not to incorporate into my stable epistemic commitments. Consider, for example, cases of perceptual illusion. I look at the line drawing, and am immediately gripped by the judgment that the top line is longer than the bottom one. This is a persistent disposition of mine; I have the same reaction time and time again, even though I have measured the lines several times and always found them to be of equal length. Still, there is no state of opinion with which I identify which renders it seriously possible that the lines do actually diverge in length; from a position of self-reflection, I recognize that I am being tricked and thus refuse to identify with my immediate judgment. Similar phenomena include irrational fears (for example, I am momentarily convinced that the small yapping dog poses a real threat to my well-being) and persistent undesirable biases (for example, I see a member of another racial group whom I know to be a solid citizen of the community, but am still momentarily gripped by the serious possibility that he will attempt to rob me). The point is the same as that made above: our concern here is with my epistemic commitments, as determined by the decisions I make from a position of reflective self-awareness, and not with my psychological dispositions to produce certain judgments given certain stimuli. The latter may be interesting to us in that they may provide some evidence about what we are committed to, but they are not to be confused with such commitments.

Levi, as one might expect, rejects the idea that a rational agent might exhibit ambivalence among multiple standards for serious possibility. He writes:

> Observe, however, that one and the same *person* may belong to different communities each of which has its own goals and values and each of which may have different and sometimes conflicting standards for serious possibility....
>
> Insofar as *X* finds himself in this predicament, he may be committed to a cognitive schizophrenia embracing different standards for serious possibility corresponding to his diverse roles....
>
> It is no doubt neither psychologically nor socially feasible to remove all inconsistency or schizophrenia. But insofar as it is feasible to do so, all agents, whether they are individual persons or communities, should commit themselves to single consistent standards for serious possibility.

> When inconsistency arises, as it inevitably will, and when it is detected, steps should be taken to eliminate it.[17]

But there does not seem to be any argument here, just dogmatism. Levi's charge of inconsistency is unfounded; the agents he describes are not committed to a single inconsistent standard for serious possibility, but are undecided between a range of incompatible standards for serious possibility. The former is clearly irrational; but there is nothing at all obviously wrong with the latter. As for his charge of schizophrenia, this seems to be nothing more than a terminological dispute: Levi chooses to pin a pejorative term on a phenomenon that I am inclined to describe more neutrally as "ambivalence." If the difference between us amounts to nothing more than this, then Levi's name calling is surely insufficient to undermine the prima facie case I have made above for the rationality of ambivalence with respect to one's possibility judgments.

At any rate, most of the disagreement between Levi and me on this point can be dissolved through a harmless shift in vocabulary. We can, if we wish, take the disjunction (union) of all of the standards for serious possibility which are reflected in my opinion; call this proposition the *minimal standard.* Those propositions entailed by the minimal standard are just those which all of the opinions I identify with regard as seriously necessary; those propositions compatible with the minimal standard are just those which are regarded as seriously possible by at least one of the opinions with which I identify; and those which are incompatible with the minimal standard are just those which I always regard as seriously impossible. So the minimal standard functions in precisely the same way as Levi's univocal standard for serious possibility. Thus, if Levi insists that rational agents are always committed to a single standard for serious possibility, we need not demur. We can simply identify his concept of serious possibility with our concept of possibility according to the minimal standard, that is, serious possibility with respect to at least one standard in the set. We can reclassify my analysis as an account of ambivalence with respect to some stronger epistemic notion, say, "quasi-serious possibility" or "contextual serious possibility."

This relabeling leaves intact a substantial point of disagreement between Levi and me, however. This is, of course, the question of the relationship between serious possibility and belief. On the account I am offering, one's beliefs will not be limited to those propositions which are entailed by the minimal standard, that is, those propositions which are regarded as seriously necessary according to all of the standards for (quasi-) serious possibility with which I identify. Reflection on cases such as the Millard Fillmore example has, I hope, convinced you that a large class of our beliefs—call them the *fallible beliefs*—are regarded as possibly false according to some of the opinions we find acceptable. (Nothing I have said is meant to rule out the possibility that one might also have *infallible beliefs*, beliefs whose truth is regarded as seriously necessary by all acceptable standards. I claim only that these beliefs do not exhaust the range of the concept under analysis.) Still, this much of Levi's view seems to be true: for any proposition I believe, that proposition must be regarded as seriously necessary according to at least one standard for serious possibility I accept. I must be able, at least in some epistemic circumstances, to maintain a judgment of total certainty in the truth of that proposition in order to count myself as believing it. Furthermore, there must not be any standard I identify with which classifies the proposition as seriously impossible. If my opinion is so ambivalent as to allow me to potentially regard a proposition as either certainly true or certainly false, then I am best described as in a state of suspended judgment with respect to that proposition. Summarizing the results of this section, we have arrived at the following necessary condition for belief:

---

[17] *The Enterprise of Knowledge*, pp. 17-18.

> *X* believes *p* only if, where *S* is the set of standards for serious possibility with which *X* identifies, *p* is seriously necessary with respect to some *s* in *S* and is seriously possible with respect to all *s* in *S*.

III

It would be nice if the necessary condition for belief we have discovered turned out to be a sufficient condition as well. If that were the case, we could pack up and go home now. Unfortunately, things do not turn out to be that simple. In the general case, it is possible sometimes to regard a proposition as seriously necessary and always regard it as seriously possible without believing that proposition.

An example: suppose that in a forthcoming election, the three front-running candidates are named Patrick Buchanan, William Clinton, and Robert Dole. These are the only candidates whose victory I take to be a serious possibility. My opinion with respect to the outcome of the election is structured as follows: in some frames of mind with which I identify, I rule out a Buchanan victory but allow for the possibility of Clinton or Dole winning. In other frames of mind, I regard a Dole victory as impossible but admit the possibility of either a Clinton or a Buchanan victory. On all of the rest of the standards that I regard as admissible, any of the three outcomes is seriously possible. (Please note that these do not represent my actual opinions with respect to any historical election.) What do I believe about the outcome of the election? Well, the proposition that Clinton or Dole will win satisfies the necessary condition for belief: it is sometimes seriously necessary and everywhere seriously possible. Likewise for the proposition that Clinton or Buchanan will win. So, if we take the necessary condition to provide a sufficient condition as well, I believe both of these propositions. But then by the requirement of deductive cogency, I am committed to believing their conjunction as well; I must believe that Clinton will win. Now, however, we have a problem. For the proposition that Clinton will win does not satisfy the necessary condition; there is no standard I identify with which judges this proposition to be seriously necessary. The assumption that the necessary condition is also sufficient has driven us into a contradiction.

What has gone wrong? What feature of my attitude toward the proposition that Clinton or Dole will win renders it unsuitable as a candidate for belief? To answer these questions, we need to take a much closer look at the role that beliefs play in deliberation and inquiry, at the reasons why we care about having beliefs at all. According to Ronald B. de Sousa,[18] we are interested in having beliefs because "we harbor epistemic desires, which make of us collectors of true sentences" (*ibid.*, pp. 58-59). On this view, beliefs can be likened to wagers that a given proposition is true: just as the aim of the gambler is to amass as many profitable bets as possible, the goal of the epistemic agent is to have as many true sentences and as few false sentences as possible in his collection. This may be true as far as it goes, but it seems to me that such a view neglects a crucial aspect of the role of belief. van Fraassen[19] comes much nearer to the mark when he writes: "The point of having beliefs is to construct a *single* (though in general incomplete) picture of what things are like...to form a single, unequivocally endorsed picture of what things are like" (*ibid.*, pp. 349-50). In other words, our beliefs are not just isolated sentences in a collection; they are meant to hang together, to tell a univocal story about the way the world is. It is this feature of belief which subjects it to the requirement of deductive cogency: we do not require the gambler to make sure that all of the propositions he bets on be logically consistent; but we do require of the storyteller that the logical consequences of what she has already said will not be contradicted as the story unfolds.

---

[18] "How to Give a Piece of Your Mind: or, The Logic of Belief and Assent," *Review of Metaphysics*, xxv (September 1971): 52-79.

[19] "Fine-grained Opinion, Probability, and the Logic of Full Belief," *Journal of Philosophical Logic*, xxiv (August 1995): 349-77.

A standard for serious possibility is in many ways like a picture or a story of the world: it represents, in effect, the strongest story about the world to which I am willing to commit myself from the perspective of a particular frame of mind. The question that faces us is: How do we derive from the various stories with which I partially identify a single story that I can be said to unequivocally endorse from a self-reflective perspective? For a restricted class of cases, the answer is trivial. These are cases in which all of the standards with which I identify are logically entailed by a single strongest standard with which I also identify. In other words: define a *maximal story* to be a standard for serious possibility with which I identify such that no other standard with which I identify is logically stronger. Then the cases I am presently considering are cases in which I identify with exactly one maximal story; and in which every other story I admit as acceptable represents a logical weakening of this maximal story. For these simple cases, there is no problem with simply identifying the belief corpus with the maximal story: the necessary condition really is sufficient in such cases. An agent who falls within this domain is, on the whole, committed to the truth of a single strong story about the world; various considerations may make him doubtful about aspects of this story and thus lead him momentarily to endorse weaker versions, but he is never moved to assent to anything that goes beyond or contradicts this single underlying picture.

This is a very tidy picture of belief. So tidy, in fact, that at one point I considered making it mandatory. In my first attempt to tackle these issues, I proposed that rational agents were subject to a normative requirement that, in effect, ensured that their opinions were governed by exactly one maximal story. I have since come to regard this move as seriously misguided. States of opinion more complicated than these simple cases now seem to me to play an important role in our epistemic activities, a role which cannot simply be swept under the rug. The election example I presented above provides a toy model of the sort of cases I have in mind: here, I identify not with one but with two maximal stories, each of which restricts the list of possible victors to two. Given the right circumstances, I can convince myself that either of these stories is certainly true. But I cannot convince myself that they are both true at the same time; and so, given that from a position of reflection I see nothing to recommend one over the other, I refrain from incorporating the specific details of either of them into my overall picture of the world. More realistic examples can perhaps be found by considering my attitudes toward various contemporary scientific theories. Sometimes, when I reflect on the tremendous predictive and technological successes of the physical sciences, I am genuinely moved, pace Peirce, to regard the central tenets of our best physical theories as certainly true. In other words, when I am in such a mood, I am inclined to paint a portrait of the world that includes the truth of these theories as a central feature. But I am also capable of identifying with much darker skeptical stories, according to which the universe is so incomprehensibly complex that we literally have no chance of doing any better than approximately capturing a few extremely local and superficial empirical regularities of no real significance. Such frames of mind are especially likely to result from prolonged exposure to studies in the history of science. It is because of this radical ambivalence that I am reluctant to describe myself as believing the results of contemporary physical science: even though there are stories of the world with which I identify according to which those results are true, I see from the broadest perspective no reason to favor those stories over other stories I identify with according to which those results are certainly false.

We are now within spitting distance of the explication we have been seeking. First, however, it is necessary to impose a new normative requirement on our epistemic commitments, one that I have previously alluded to but not yet made explicit. In the preceding discussion, I have generally taken it for granted that an agent will be committed to at least one maximal story. Nothing I have said so far guarantees this, however; it is possible to construct coherent sets of possibility standards in which no story is maximal. We can imagine, for example, an infinite set of standards, linearly ordered by logical implication (that is, given any two standards in the set, one of them implies the other), which converge upon but never reach a lower bound which is not in the set. Such cases, though imaginable, must be ruled out if the analysis I am proposing is to succeed. Specifically, what is required is the following:

> *Boundedness condition*: let S be an agent's set of standards for serious possibility; for all s in S, there is a maximal story m in S such that m logically implies s.

In other words, for every story about the world that an agent is willing to tell on some available frame of mind, it is possible to supplement that story with additional details and thus construct a maximal story that the agent is also willing to tell on some other available frame of mind. I am not claiming that this normative requirement recommends itself to us with a great deal of intuitive force, nor that it plays a significant role in our ordinary way of thinking about our epistemic practices. (I must admit that, purely in my own case, the principle does carry with it at least some intuitive force. There does seem to me to be something genuinely wrong about states of opinion that violate boundedness, such as the converging series described above. But my intuitions on the matter are hardly neutral or untutored.) My argument for imposing this principle is that it greatly simplifies our understanding of the nature of belief without imposing undue strain on our other epistemic concepts. It does not seem to me that, in ruling out unbounded epistemic states, we are arbitrarily dismissing a class of opinions that are of genuine epistemological significance (as we would have done had we ruled out opinions governed by multiple maximal stories). Any unbounded opinion can easily be transformed into a bounded one simply by adding the appropriate maximal stories to the standard set; this modification does not seem to produce unacceptable changes in the underlying state of opinion thereby represented. Given this, I think that the theoretical advantages to be gained by imposing the boundedness requirement are sufficient to justify its adoption. (Still, I do regard this move as an unfortunate necessity; and I would be deeply obliged to anyone who can show me a simple way to reconstruct the heart of my analysis in such a way as to avoid this requirement.)

So, for agents who satisfy the boundedness condition, I propose the following definition:

> An agent's belief corpus is the disjunction (union) of all of the maximal stories with which that agent identifies.

As before, the agent believes just those propositions which are logically implied by the corpus. This analysis trades on the insight that from the self-transcendent perspective relevant to the determination of one's overall epistemic commitments, one is not in a position to choose between the various precise states of opinion with which one partially identifies. *Ex hypothesi,* all of these states strike one's reflective self as of roughly equal acceptability. Therefore, in formulating a single picture of the world to which one is unequivocally committed, one should not include features that are idiosyncratic components of some but not all of the maximal stories with which one identifies. Rather, one should restrict one's attention to those claims which are common to all of the acceptable maximal stories. It is an easy matter to verify that, on this explication, all of an agent's beliefs satisfy the necessary condition formulated in section II. For any believed proposition *p* is a logical consequence of the corpus, which is in turn a logical consequence of each of the maximal stories; since the maximal stories are sometimes judged to be seriously necessary along with their logical consequences, *p* is sometimes judged to be seriously necessary. Likewise, since every accepted standard for serious possibility is implied by some maximal story, every accepted standard is logically compatible with all of the consequences of at least one maximal story. Since *p* is a consequence of every maximal story, *p* is thus judged to be seriously possible by every accepted standard. It is also trivial to see that beliefs on this model are deductively cogent. So the proposed analysis is at least formally adequate. It also seems to be intuitively acceptable, at least for all the cases we have considered here. Consider, for example, the election case once more: the explication yields the result that I believe nothing stronger about the election than that one of the three front runners will win, and this seems to be the appropriate conclusion to draw.

It is perhaps worth mentioning that this component of my analysis shares certain structural similarities with an account of belief recently proposed by Mark Kaplan.[39] Kaplan, like me, takes the project of belief formation to be the project of constructing a single univocal account of how things are, or, as he puts it, a *global theory of the world*. Kaplan calls a proposition a *global contender* if it is unsurpassed by any other proposition according to standards of plausibility, explanatory and question-answering power, comprehensiveness, and so forth. Roughly speaking, there is no way of improving a global contender along one of these evaluative dimensions without sacrificing some of the virtues it possesses among the other dimensions. (Kaplan's definition is, of course, more precise than this, but the details need not concern us.) Kaplan's hypothesis is that each global contender represents for you an equally, maximally good candidate for your global theory of the world. Since you have no further epistemic standard by which to rank the global contenders, you ought not arbitrarily favor one over another, but should rather assent to only those claims which are common to all global contenders. Hence, your global theory of the world will be the strongest proposition that is implied by all of the propositions you regard as global contenders; your beliefs will be just those propositions implied by your global theory of the world. The overlap in our perspectives should be obvious. We differ, of course, in our accounts of what it is to judge a proposition to be a global contender, that is, a maximal story.

---

[39] *Decision Theory as Philosophy* (New York: Cambridge, 1996).

I want to end this section by returning to the question I posed earlier: What is the function which beliefs play in our practices of deliberation and inquiry? Levi makes much of this question; indeed, he issues it as a straightforward challenge to those who, like me, reject his account of belief:

> Denying this thesis of *epistemological infallibilism* entails rejection of the view of knowledge as a standard for serious possibility and, hence, the view of how knowledge functions as a resource for inquiry and deliberation which I am advocating. Those who reject epistemological infallibilism, therefore, are under some obligation to supply an alternative view of the functions and value of knowledge.[22]

This challenge may seem especially pressing given the somewhat baroque analysis of belief I have been pressing. It may not be at all obvious to you why you should care whether a given proposition is or is not entailed by the disjunction of all the maximal stories you identify with. Thus, if I have successfully convinced you that you believe just those propositions to which you bear this relation, you may find yourself asking why the question of what one believes matters at all. I shall not attempt to offer a full answer to this question, but I think the above discussion provides the foundation for at least a partial answer that should be sufficient to allay any deep worries about the relevance of belief.

Belief, I claim, bears a special relationship to assertion. Suppose that I am participating in a highly idealized form of public inquiry in which my only motivation is to share the truth as I see it with my fellow inquirers. More specifically, my intention is to assert precisely those propositions which I am presently certain are true. It is taken for granted that my opinion does not change within the context of this inquiry; that is, all of my assertions will be taken as reflections of a single underlying state of opinion (although psychological factors such as salience may well shift my perspective on this stable opinion in the course of deliberation). Granted, real-life inquiry almost never takes this pure form, but this fact does not negate the philosophical significance of the example. I ask myself: By what standard should I judge what to assert and what not to assert? One possible answer: I should assert just those propositions which presently strike me as seriously necessary. Hence, if I am in one of my skeptical moods, I shall assert very little; if my outlook changes so that I am more inclined to assent to a more detailed picture of the world, I shall become more garrulous. At the limit, I may be moved to assert one of my maximal stories in its entirety. Another possible answer: I should assert just those propositions which presently strike me as seriously necessary *and which I believe*. Any proposition that goes beyond my belief corpus, even if I presently find it completely compelling, should not be introduced into the public debate. My claim is that the second answer is the better one. Belief in a proposition is a necessary prerequisite for warranted assertion of that proposition (at least when one's motives are wholly truth governed).

---

[22] *The Enterprise of Knowledge*, p. 13.

The argument for this claim relies, once again, upon the condition of deductive cogency. The other participants in the conversation do not regard my assertions in strict isolation from one another; they assume that my assertions are meant to hang together as contributions to an overall story about the world which I am in the process of laying out. They will therefore take it for granted that I shall not make any contradictory assertions, and that I am committed to the truth of any consequence they can derive from the collection of assertions I have made thus far. The presupposition of deductive cogency is a basic component of all serious epistemic inquiry. But if I take my immediate standard of serious possibility as my criterion for assertion, I shall not in general be upholding my commitment to deductive cogency. Once more, the election case provides a good example: if, at one stage, I am moved to assert that either Clinton or Dole will win, and then at a later stage, I am driven to assert that Clinton or Buchanan will win, my companions will be justified in regarding this pair of assertions as tantamount to an assertion on my part that Clinton will win. But I do not regard myself as warranted in asserting this proposition; hence, I cannot in good faith permit myself to make a set of assertions that can be misinterpreted in this way.

Here is where my belief corpus comes into play. This corpus, recall, represents the strongest proposition that is logically implied by all of the maximal stories to which I am committed. Hence, if I restrict my assertions to those propositions which I believe, I need never worry about committing violations of deductive cogency of the sort described above. I shall never make an assertion whose truth I shall later be moved to rule out as a serious possibility; nor shall I make any set of assertions that jointly entail a proposition I am under no circumstances inclined to regard as certainly true. In other words, as long as I take my belief corpus to provide the standard for assertibility, I shall ensure that I am always in a position to assert as large a chunk of my presently favored maximal story as possible, subject to the constraint that I must pay due respect to the other maximal stories toward which I may soon find myself leaning. This, I submit, provides sufficient reason in and of itself for us to care about what we believe. By keeping track of our beliefs, we are in effect keeping track of the limits of what we may safely assert (or, mutatis mutandis, take as a stable premise in our private deliberations) without fear of contradicting ourselves.

JONATHAN ROORDA[*]

---

[*] Jonathan Roorda was a graduate student in the Department of Philosophy at Princeton University at the time of his death of cancer in June 1996. He was a few months away from completing a Ph.D. dissertation on a probabilistic explication of full belief. He was twenty-five. This article was prepared for publication by his advisor, Richard C. Jeffrey.

BOOK REVIEWS

The Metaphysics of Free Will: An Essay on Control. JOHN MARTIN FISCHER. Cambridge: Basil Blackwell, 1994. ix + 273 p. Cloth $52.95, paper $21.95.

It seems a truism that our status as persons, beings with a special moral status as bearers of rights and (above all) responsibilities, is connected with our having control over what we do. But how are we to understand the notion of control that figures in this truistic thought? In his meticulous and illuminating study, *The Metaphysics of Free Will*, John Martin Fischer distinguishes between two different kinds of control that personhood might involve. *Regulative control* is the ability to determine what one does by selecting the action one is to perform from a range of genuine alternatives; persons only have control over what they do, in this regulative sense, to the extent it is true of them that they could also do otherwise. *Guidance control*, by contrast, is not in the first instance a matter of the alternatives that are open to one in action, but of what one actually does. One exhibits guidance control when one's actions result from a process that is appropriately responsive to reasons, and this condition seems satisfiable even in cases in which one could not, in fact, do otherwise.

The bulk of Fischer's book (chapters 1-6) is a defense of the claim that regulative control is not compatible with determinism or with the existence and omniscience of god. In the final three chapters, however, Fischer changes tack, arguing that the challenge posed by these incompatibilist arguments can be laid to rest, because regulative control is not really a genuine condition of responsibility after all; guidance control alone will suffice. The resulting position is characterized as a form of "semi-compatibilism" (180): moral responsibility is compatible with determinism and god's existence, since it only requires guidance control, while freedom to do otherwise (regulative control) is not so compatible. To my mind, this position exhibits a curious instability, in that the first of its constituent claims undermines the interest of the second. If, as Fischer thinks, moral accountability only requires guidance control, the thesis that there is a different kind of control that is not compatible with determinism begins to acquire a scholastic air. Why should we care?

Fischer's defense of incompatibilism about regulative control is nevertheless highly sophisticated. He begins (chapter 1) by laying out the kind of incompatibilist argument familiar from the work of

0022-362X/97/9403/156–59

Carl Ginet, Peter van Inwagen, and David Wiggins, highlighting in particular the role of the "principle of transfer of powerlessness" in the argument. Challenges to this "transfer principle" from Anthony Kenny and Michael Slote are then parried (chapter 2); and it is argued that the principle by itself—independent, that is, from the truth of determinism or the existence of an omniscient divinity—is not sufficient to call in question the idea that we have regulative control (chapter 3). Neither is it necessary, however, for Fischer argues that a skeptical challenge to regulative control can be mounted even without the transfer principle (chapter 3). In particular, he develops a "conditional" version of this challenge, which gives expression to deeply appealing assumptions about the fixity of the past, the fixity of the laws of nature, and the omniscience of god (62-63).

As those familiar with recent debates about free will are aware, however, these assumptions are themselves subject to various interpretations, not all of which lend support to an incompatibilist treatment of regulative control. In chapter 4, Fischer explores different ways of understanding the fixity of the past and the laws. He presents thoughtful versions of "local miracle" and "multiple pasts" compatibilism, according to which the intuitive force of the fixity assumptions is preserved without abandoning the idea that agents have the power to do otherwise, even in a deterministic world. These positions cannot be dismissed as merely ad hoc responses to the threat of incompatibilism; but neither can conclusive considerations be advanced in their favor, in light of the continuing independent appeal of stronger, incompatibilist interpretations of the fixity assumptions. Fischer concludes—persuasively, in my view—that the debate about the conditional argument for incompatibilism has reached a "dialectical stalemate" (83-85).

This is not the end of the matter, however, for Fischer believes the argument for incompatibilism about regulative control can be reformulated in terms that are at once simpler and more compelling. He presents a "basic" version of incompatibilism in chapter 5, sketching its implications for Newcomb's problem, and he explores some of the presuppositions of this argument (specifically about the fixity of the past) in chapter 6. The basic version turns on the idea that we only have it in our power to do those things which are extensions of the actual past, holding fixed the laws of nature (88). As Fischer himself recognizes (see 199), the basic version is probably better described as a picture than as an argument. His interesting and surprising suggestion seems to be that articulation of this picture can lead to a restructuring of the debate, in which the power and appeal of the incompatibilist position become fully apparent for the first time.

I have my doubts, however, about how much is achieved by this move. Fischer contends that, with the help of the basic picture, we shall be able to distinguish between the truth conditions of 'can' claims and of apparently related counterfactual conditionals that associate "local miracles" or alternative pasts with an agent's exercise of the ability to do otherwise; this, in turn, undermines the compatibilist analyses of 'can' claims in terms of such conditionals (89-94). But the crucial question is why we should accept the incompatibilist interpretation of 'can' that motivates this distinction of senses. Fischer himself points to the "natural appeal" (89) of the basic picture, and to the light it can shed on Newcomb's problem. But given the sharp controversy that surrounds the latter problem, Fischer's treatment of it seems unlikely to enhance the attractiveness of the basic picture for those not already in its grip. As to the inherent appeal of the basic picture, it could plausibly be retorted that what seems attractive about it are just the familiar ideas of the fixity of the past and of the laws, and that these ideas can be preserved on a different, compatibilist interpretation of 'can'. We seem to be back to the dialectical stalemate.

If anything is likely to restructure the debate, it is Fischer's argument, in chapter 7, that regulative control is not anyway a condition of moral responsibility. The argument builds on a well-known form of counterexample to the principle of alternative possibilities, due to Harry Frankfurt, in which circumstances (for example, the watchful presence of a "nefarious neurosurgeon") ensure that an agent is deprived of alternative possibilities, without contributing to explain what the agent actually does. The success of these counterexamples has been questioned by a number of philosophers, who point out that the agents in the counterexample situations seem to retain some residue of freedom to do otherwise, despite initial appearances to the contrary. But Fischer rejects this "flicker of freedom" strategy, contending that the alternatives it postulates are not "robust" enough to ground our responsibility ascriptions (140-47). In defense of the flicker of freedom strategy, however, it might be suggested that it is not the robustness of the alternatives postulated which is to ground responsibility, but just the presence of guidance control by the agent. The point of the strategy might lie in the difficulty of describing a clear case in which this guidance condition is satisfied without the agent's also having some alternative possibilities; this, in turn, may undermine the initial assumption that guidance control is really attainable in the complete absence of such alternatives.

Fischer presents his own account of guidance control in chapter 8. On this account, moral responsibility requires that the "actual sequence" leading to action involve a reasons-responsive mechanism; a mechanism is appropriately (that is, "weakly") reasons responsive, in turn, if there is some possible world in which the agent has sufficient reason to do otherwise, the mechanism operates, and the agent does otherwise (164-68). I have elsewhere[1] expressed some reservations about this account—it has the implausible consequence, for instance, that young children and the mentally deranged are morally accountable (as Fischer concedes, 243). Fischer plans to refine his theory to meet these and other objections in future work (206-07); but there are basic difficulties inherent in his talk about "reasons-responsive mechanisms" that make me skeptical about the prospects for success along this route. Take the idea of reasons responsiveness. By Fischer's criteria, a mechanism will exhibit this characteristic if it leads an agent to do otherwise in some possible world in which there is sufficient reason so to act. But this analysis leaves out something that seems essential to the idea of *responsiveness* to reasons, namely, that an agent should act in a certain way *because* there is sufficient reason to do so (compare Fischer's own initial presentation of reasons responsiveness (164)). The peculiar consequence is that Fischer is willing to classify as reasons responsive even physical processes (174) or the rogue desires and mental tropisms involved in weakness of will (166).

The deeper problem, however, lies in the supposition that questions of moral accountability can be clarified by attending exclusively to the modal properties of the "mechanisms" involved in action. This approach brings an objectifying, third-personal vocabulary to bear on phenomena that have their natural place within the deliberative perspective of practical reason, with the result that the intuitive locus of responsibility, the person, seems to drop out of view. A better approach would focus not on the counterfactual behavior of the *mechanisms* that issue in action, but on the normative competence of the *agents* who perform those actions, and are morally accountable for them.

R. JAY WALLACE

Humboldt-Universität zu Berlin

---

[1] In *Responsibility and the Moral Sentiments* (Cambridge: Harvard, 1994), pp. 189-90.

BOOK REVIEWS

From a Biological Point of View: Essays in Evolutionary Philosophy. ELLIOTT SOBER. New York: Cambridge University Press, 1994. x + 255 p. Cloth $65.00, paper $19.95.

This is a collection of essays by a leading philosopher of biology, on topics which for most part are outside of the philosophy of biology. As the title suggests, the book brings a "biological point of view" to a set of fairly traditional topics in the philosophy of science, the philosophy of mind, and ethics. As Elliot Sober notes, this biological point of view contrasts with the logical point of view that has dominated twentieth-century responses to some of these topics, particularly those in the philosophy of science. Sober says in his introduction that his title is intended both to pay "homage" to, and to contrast with, the famous 1953 collection of essays in which W. V. Quine[1] exemplified this logical point of view.

The essays in the first half of the book or so (chapters 1-5) deal with topics in the philosophy of mind, moral psychology, and ethics. Most of these essays appear here for the first time. The essays in the second part (chapters 6-12) are mostly concerned with issues in the philosophy of science, such as confirmation, explanation, and causality. These essays have all been published before, and most appeared fairly recently in major journals.

A number of the essays in the second part discuss themes that Sober developed in a more specialized context in his excellent 1988 book *Reconstructing the Past.*[2] In that work Sober argued that principles of simplicity, parsimony, and economy (such as "Occam's razor") only have a legitimate role to play in theory choice when there is a specific model telling us why and how simplicity is to be expected over complexity in that context. For example, we should prefer one evolutionary hypothesis over another on the basis of simplicity only if we have a model of how evolution works which tells us that simple evolutionary processes occur more often, in the domain in question, than complex ones. There is no wholly *general* justification for preferring simplicity.

I divided the book into a philosophy of mind part and a philosophy of science part, but there are various connections between the

---

[1] *From a Logical Point of View* (Cambridge: Harvard, 1953).

[2] Cambridge: MIT, 1988.

0022-362X/97/9403/160–64

two sets of themes. Sober's view about parsimony is applied, briefly but effectively, in a discussion of psychological egoism (24). Sober describes some simple models of the evolution of altruistic behavior and finds no reason to say that evolution will tend to produce psychological egoists. He notes that social scientists often opt for a psychological egoist view because it is supposedly "more parsimonious" than a hypothesis of psychological altruism. But, Sober claims, in the absence of a specific model that tells us that psychological egoism is actually more likely to be produced, there is no justification for this preference.

There are some very good essays in the collection, but considered as a whole the book does have a significant shortcoming. It needed a substantial introduction, both to tie the chapters together and to address some questions that arise repeatedly within various chapters. The book needed an introduction of about twenty pages, but Sober gives us only seven. In this short essay Sober warns the reader against looking for too much unity in the essays, and against erecting a monolithic "biological point of view" to replace earlier monoliths. Perhaps that explains the brevity. But a longer discussion, even of this disunity, would have been welcome.

The biggest issue left unresolved concerns the role of Bayes's theorem in the confirmation of theories. Sober uses Bayes's theorem in several of the essays (chapters 6, 7, and 8), but he always accompanies the theorem with expressions of ambivalence about its status. He denies that, in general, there are objective prior probabilities associated with theoretical hypotheses, although he does think that the likelihood terms (probabilities of observations conditional upon the truth of hypotheses) can be interpreted objectively. If there are no objective prior probabilities, it is hard to view Bayes's theorem as describing the real structure of the confirmation relation unless one is a subjectivist, which Sober is not. Sober's brow appears frequently furrowed by this problem. But Sober continues to *use* Bayes's theorem, claiming that it has heuristic usefulness nonetheless. Bayes's theorem is a "convenient reminder" that the evaluation of hypotheses must take into account the antecedent "plausibility" of a theory as well as how it deals with evidence (139); Bayes's theorem is useful in "classifying" the considerations that affect a theory's believability (137). But I do not think so many readers need a "reminder" of this type. Bayes's theorem might remind and classify, but it also appears to assert a *specific mathematical relationship* between the factors that affect confirmation. This is what Sober remains ambivalent about, and it needed some discussion in the introduction.

There are also interesting unresolved issues concerning realism. In chapter 6, "Contrastive Empiricism," Sober argues for an empiricist view, not about untestable *propositions*, but about certain *problems*: if two hypotheses are empirically equivalent, science cannot discriminate between them with respect to their believability. When a hypothesis is compared to one that is not empirically equivalent to it, a scientific choice can be made and there is no need for empiricist scruples about what the theories assert. But if the same hypothesis is compared to an empirically equivalent one, there can be no rational choice between them, however much we find ourselves preferring one over the other (129). Empirical equivalence is understood just as bestowing the same likelihood on all possible observations. So this is a very austere framework for discussing theory choice.

If this position is adopted, a problem arises with the status of Darwinism and the "biological point of view." As I understand his project, Sober aims to take a Darwinian viewpoint seriously and work out what we should expect, in the light of this, for questions concerning altruism, mental content, and so on. I understand this project as requiring that we see Darwinism as a *factually accurate description* of our history as a species. We admit that Darwinism is fallible, but view it as well confirmed and something we can reasonably take to be true. Yet Sober also defends a view in which the choice between Darwinism and a suitably tailored creationist theory—one that predicts all the same observations as Darwinism—is not scientifically decidable. (I assume here that the relevant possible observations are present and future ones (see 127). If possible past observations are included, the creationist has a tougher time.) The essay on empiricism is not specifically about Darwinism, and the status of evolutionary theory is not discussed there. That is not a problem when the essay is considered in isolation. But the issue is important given the aims of other chapters.

In "Contrastive Empiricism," Sober is also surprisingly easy-going about the position to which he finds his arguments pointing—a combination of naturalism with skepticism (129). I do not see this combination as self-refuting, but it is certainly an uncomfortable place to end up—and not an advertisement for the biological point of view.

I shall finish with some discussion of the essay I found most interesting of all, "Apportioning Causal Responsibility" (chapter 10). Here, Sober discusses the question of when and how we can say, of two causal factors implicated in some effect, that one factor played *more* of a role than the other. In particular, Sober discusses the problem of apportioning causal role between genes and environment, when explaining a structural or behavioral ("phenotypic") property of an individual. The trend in philosophy of biology has been to argue that, given the nature of the interactions between genes and environment, there is no sense in trying to decide which made "more" of a contribution to, say, an individual's height. Although the statistical method called "analysis of variance," or ANOVA, appears to generate measurements of the relative contributions of these factors, influential articles by R. Lewontin[3] and others have attacked the idea that ANOVA gives us causal information of this type. In particular, it is often claimed that ANOVA statistics give no information about individual cases—they describe only population-level properties.

Sober argues that ANOVA methods do provide a basis for claims about the relative causal roles of genes and environment. But these claims are of a specific type. We cannot say that genes played more of a role in bringing about some individual's height than her environment did. We could only make that sort of claim if genes and environment exerted their effects via some common currency, which they do not. But we can, with ANOVA, decide which factor *made more of a difference* to her height.

An ANOVA describes patterns of correlation, in a given population, between environments and phenotypic values, and between genotypes and phenotypic values. Sober's claim is that these population statistics provide a basis for certain counterfactuals. They make it possible to say that, if Jane had had the same genotype but a different environment, the result would have been such-and-such. An analogous counterfactual can be asserted about what would have happened if she had had the same environment but a different genotype. Comparing the two, we can decide which factor, genes or environment, made more of a difference to her height. So Sober's strategy is to take the information about the entire population which an ANOVA provides, and view this information as grounding counterfactuals about the individual, Jane.

Sober makes this case in an interesting way, and many issues are raised. I shall not discuss the issue of whether the causal claims in question are properly analyzed with these counterfactuals. The point I shall make is that we should not think that ANOVA statistics make the *status* of these counterfactuals less philosophically problematic than usual. Beneath this machinery is just the idea that because *other* individuals in the population have certain genotypes, environments, and phenotypes, and have them in certain frequencies, it is true *of Jane* that if she had had a different genotype the results would have been a certain way. Sober concedes in footnotes that counterfactuals can sometimes be vague, indeterminate, and so on, but when the chips are down (as in his final paragraph) he does claim that it is a factual matter which factor made more of a difference to Jane. It is hard, however, to see these modal facts as *given* to us by the biology or by the statistics. Sober says that it is "natural" to make certain counterfactual claims about individuals on the basis of facts about the population (191); that the context "allows us to fill out" the counterfactuals (197). The counterfactuals may indeed feel "natural" as we say them, but they are still a controversial addition to the mathematical description of Jane and her fellows.

---

[3] On interpreting ANOVA, see Lewontin, "The Analysis of Variance and the Analysis of Causes," reprinted in R. Levins and Lewontin, *The Dialectical Biologist* (Cambridge: Harvard, 1985). Sober's presentation of some of the mathematical details of ANOVA is unorthodox and, in my view, misleading. The overall argument of the paper is not affected by this.

This point about counterfactuals is reminiscent of what I said above about Bayes's theorem. An apparently helpful way of talking may nonetheless involve problematic metaphysical commitments.

Readers seeking a detailed program to guide future work, a systematic "biological point of view," will not find it here. But they will find a wide-ranging collection of adventurous essays, which challenge commonly-held views about science and justification in deceptively simple and often compelling ways. Like the rest of Sober's work, the whole collection is written with exceptional clarity.

PETER GODFREY-SMITH

Stanford University

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NO. 4, APRIL 1997

THE EPISTEMIC STATUS OF PROBABILISTIC PROOF*

> Proof is always a relative thing. It's an overwhelming balance of probabilities.
> —Philip Marlowe

A deductive argument is considered by mathematicians (and many philosophers) to be the *only* legitimate way to establish that a mathematical claim is true. I shall argue that mathematicians do not have good grounds for their rejection of probabilistic methods as a means of establishing mathematical truths. Specifically, I shall argue that one particular probabilistic method (which utilizes recent DNA technology) has no *epistemic* drawbacks that are not shared by standard methods of establishing mathematical truths.

Mathematicians are primarily engaged in establishing the truth of mathematical claims. They typically establish that a mathematical claim is true by writing down a proof. Of course, writing down a proof is not the only method of establishing the truth of a mathematical claim that mathematicians find acceptable. For example, it is acceptable to use a computer to establish the truth of a mathematical claim.[1]

Nevertheless, there are some methods that are clearly *not acceptable* to mathematicians. For instance, the mathematician Leonhard Euler[2] says that "we should take great care not to accept as true such proper-

---

* I would like to thank Jody Azzouni, Mark Balaguer, Jeff Barrett, John Lewis, Peter Lewis, Penelope Maddy, Kay Mathiesen, Carl Pomerance, Amie Thomasson, Peter Woodruff, and Dan Zelinski for reading and commenting on this work at various stages in its development.

[1] On the proof of the four-color theorem, see, for example, Ian Stewart, *The Problems of Mathematics* (New York: Oxford, 1987), pp. 116-18.

[2] Quoted in G. Polya, *Mathematics and Plausible Reasoning, Volume I* (Princeton: University Press, 1954).

0022-362X/97/9404/165–86

ties of the numbers which we have discovered by observation and which are supported by induction alone" (*ibid.*, p. 3). In a similar vein, the computer scientist David Harel[3] claims that "as long as we use probabilistic algorithms only for petty, down-to-earth matters such as wealth, health, and survival, we can easily make do with very-likely-to-be-correct answers to our questions. The same, it seems, cannot be said for our quest for absolute mathematical truths" (*ibid.*, p. 332).

I shall argue that mathematicians do not have good grounds for this across-the-board rejection of probabilistic methods *as a means of establishing mathematical truths.* Toward this end, I shall compare one particular probabilistic method (which I refer to as *probabilistic DNA proof* (PDP)) with several methods (such as writing down a proof or giving a computer proof) that mathematicians consider to be acceptable means of establishing mathematical truths.[4] I shall argue that there is no epistemically important qualitative difference between PDP and the methods acceptable to mathematicians. Specifically, I shall argue that every quality that might be suggested as an epistemically important difference falls under at least one of the following three categories:

Category A: There are methods acceptable to mathematicians which do not have the quality.
Category B: PDP does have the quality.
Category C: The quality is not epistemically important (that is, it is not important to the job of establishing mathematical truths).

Before beginning, I want to mention a few claims for which I am not arguing. First, I am not *merely* asserting that probabilistic methods can provide good reason to believe that a mathematical claim is true. Mathematicians (including Euler and Harel) seem to be willing to concede this point. Second, I am not claiming that *every* probabilistic method can legitimately be used to establish that a mathematical claim is true. Finally, it does not follow from my thesis that mathematical practice needs to be revised. In fact, mathematicians probably do have good reasons (for example, sociological and/or pedagogical ones) for not using probabilistic methods. I am only claiming that they do not have good *epistemic* reasons.

[3] *Algorithmics: The Spirit of Computing* (Reading, MA: Addison-Wesley, 1987).

[4] I shall not stack the deck by comparing PDP with methods where there is a serious controversy over whether or not a mathematical truth has been established (such as Wu-Yi Hsiang's proof of Johannes Kepler's conjecture about packing spheres; see Thomas C. Hales, "The Status of the Kepler Conjecture," *Mathematical Intelligencer*, XVI, 3 (1994): 47-58).

In section I, I shall explain how PDP works. In sections II-IV, I shall consider three main types of objections to my thesis.[5] First, it might be suggested that PDP does not provide enough certainty that a mathematical claim is true. Second, it might be suggested that PDP does not give us a priori knowledge that a mathematical claim is true. Third, it might be suggested that PDP does not provide a proof that a mathematical claim is true.

I. THE PROBABILISTIC METHOD

The computer scientist Leonard Adleman[6] has recently discovered a way to use DNA technology to solve difficult combinatorial problems. The particular probabilistic method that I shall discuss is based on Adleman's work and is a method of establishing that there is no *Hamiltonian path* (HP) in a given directed graph.

A *directed graph* can be thought of as a group of towns connected in various ways by one-way roads. A *path* is a sequence of towns such that each town in the sequence is connected by a one-way road to the next town in the sequence. An HP is a path that visits every town in the graph exactly one time.

There is a fairly simple deterministic algorithm that can be used to establish that there is no HP in a given directed graph (containing *N* towns). We simply check (in an orderly fashion) every possible sequence of *N* towns to see if it is an HP. If we exhaust the large, but finite, list of possible sequences without finding an HP, then there is no HP in the graph. If a mathematician actually uses a digital computer to perform this algorithm, then we shall say that she has given *deterministic computer proof* (DCP) that there is no HP in the graph.

Unfortunately, as the number of towns (*N*) increases, the number of possible sequences of *N* towns increases exponentially. As a result, if a graph contains a very large number of towns, it is too time consuming to use DCP. It is feasible, however, to perform an algorithm that checks the possible sequences all at once. DNA technology allows us to perform such an algorithm.

As is well known, DNA is a powerful device for encoding large amounts of information; a single strand of DNA can encode all of the information necessary to generate a human being. As I shall explain below, we can also use strands of DNA to represent the towns and roads in a directed graph and the paths through the graph. Furthermore, we can perform the following algorithm by manipulating these strands of DNA:

---

[5] A short summary of a few of these arguments appears in my "Mathematical Proof and the Reliability of DNA Evidence," *American Mathematical Monthly*, CIII, 6 (1996): 491-97.

[6] "Molecular Computation of Solutions to Combinatorial Problems," *Science*, CCLXVI (1994): 1021-24. Adleman actually performed the algorithm described below (with some slight differences) on a directed graph containing seven towns.

Step 1: Generate a whole bunch of paths through the graph.
Step 2: Remove all paths except those which visit exactly *N* towns.
Step 3: Remove all paths except those which visit every town.

Together, step 2 and step 3 eliminate any path through the graph which is not an HP. If a mathematician performs these three steps and no paths (that is, strands of DNA) remain, then we shall say that she has given PDP that there is no HP in the graph.

Of course, it is possible for this algorithm to give the wrong answer. It is not necessarily the case that all paths that visit exactly *N* towns are generated in step 1. As a result, it might happen that there are HPs in the graph, but that no strands of DNA representing HPs remain after the three steps are performed because no such strands were generated in step 1. In the remainder of this section, however, I shall explain precisely how DNA technology can be used to perform this algorithm and why the probability is very low that this algorithm will give the wrong answer.

A single strand of DNA is a chain of simpler molecules called *bases*. The four types of bases are designated *A*, *T*, *G*, and *C*. So for example, *CTTGAG* represents a 6-molecule long strand of DNA. Under the right circumstances, two strands of DNA will bond together to form the familiar double helix configuration. Since the base *A* will bond only with the base *T* and *G* will bond only with *C*, however, the two strands have to be the inverses of each other (where *A* is the inverse of *T* and *G* is the inverse of *C*). So, for example, two 6-molecule long strands of DNA might bond in the following configuration:

CTTGAG
| | | | | |
GAACTC

In the first step of the algorithm, each town in the graph is represented by a (randomly chosen) 20-molecule long strand of DNA (for example, *GCTATTCGAGCTTAAAGCTA*). Thus, a path (that is, a sequence of towns) can be represented by several town strands linked together end to end. Huge quantities of each town strand are placed in a test tube. The problem is to get them to link together to form strands that represent paths.

This problem is solved by adding huge quantities of each road strand to the test tube. Each road in the graph is also represented by a 20-molecule long strand. Each road strand is designed so that it will link together the two strands that represent the two towns that the road connects. For instance, suppose that there is a road going from town *X* to town *Y*. The first 10 molecules of this road strand are the inverses of the last 10 molecules in the town *X* strand; the last 10 molecules are the inverses of the first 10 molecules in the town *Y* strand (see figure 1 below). Thus, this road strand can bond with the second half of the town *X* strand and the first half of the town *Y* strand. As a result, this road strand can serve as "a molecular splint" to hold the town *X* strand and the town *Y* strand together.[7]

```
{         Town X         }{         Town Y         }
GCTATTCGAGCTTAAAGCTAGGCTAGGTACCAGCATGCTT
          ||||||||||||||||||||
          GAATTTCGATCCGATCCATG
          {  Road from X to Y }
```

Figure 1

The town strands and road strands are then allowed to bond together to form double helices. One strand of each double helix will consist of town strands linked together end to end representing a path through the graph. To complete the first step of the algorithm, a ligation reaction glues the town strands together so that they will not come apart even when the molecular splints (that is, the road strands) are removed.

In the second step, a gel electrophoresis separation process sorts the strands of DNA by length. Only those strands which are 20 x *N* molecules long are retained. In other words, only those strands which represent paths that visit exactly *N* towns are retained.

In the third step, the test tube is heated until the double helices break apart into single strands of DNA. Next, magnetized copies of the inverse of the strand that represents the first town in the graph are added to the test tube. The magnetized inverses are then allowed to bond with the single strands of DNA in the test tube. The magnetized inverses will bond only with strands that represent paths that visit the first town. Next, a magnet is held against the test tube and these paths (that is, those which visit the first town) are retained while the rest are poured away.[8] This procedure is repeated for each town in the graph. At the conclusion, the test tube contains only paths that visit every town in the graph.

---

[7] See Keith Devlin, "Test Tube Computing with DNA," *Math Horizons* (April 1995): 14-21, here p. 17.

[8] At this point, the test tube is again heated in order to separate and remove the magnetized inverses.

As I noted above, it is possible for this algorithm to give the wrong answer. In other words, it might happen that there are HPs in the graph, but that no strands of DNA representing HPs remain in the test tube after the three steps are performed because no such strands were generated in step 1. Nevertheless, the probability is very low that this algorithm will give the wrong answer because the probability is very high that all paths that visit exactly *N* towns will be generated in step 1. As I noted above, huge quantities of each town strand and each road strand are placed in the test tube in step 1. According to Adleman, "the quantity used should be just sufficient to ensure that during the ligation step (step 1) a molecule encoding a Hamiltonian path will be formed with high probability if such a path exists in the graph" (*op. cit.*, p. 1023). As a result, if no strands of DNA representing HPs are generated in step 1, then the probability that there are no HPs in the graph is very high. According to the mathematician Keith Devlin, "for all practical purposes this result could be taken as certainty, since most 'definite' conclusions in everyday life are based on far lower probabilities" (*op. cit.*, p. 16).

Before moving on to the three main types of objections, I want to discuss quickly a few obvious objections to my thesis. First, it might be suggested that the problem with PDP is that it does not provide us with an understanding of why a theorem is true. According to Ludwig Wittgenstein,[9] "proof, one might say, does not merely show *that* it is like this, but *how* it is like this" (*ibid.*, p. 159). It is apparent, however, that while providing understanding is nice, it is not required. For instance, DCP does not provide much understanding of why there is no HP in a given directed graph, but it is an acceptable method.

Second, it might be pointed out that the fact that there are no HPs in a given directed graph is not what mathematicians would typically call a theorem. According to G. H. Hardy,[10] despite "the extreme speciality of both the enunciations and the proofs" (*ibid.*, p. 105) of such combinatorial problems, they are nevertheless "genuine mathematics" (*ibid.*, p. 88) and presumably subject to the same standards of proof. Also, it should be noted that the proofs of many important theorems (such as the four-color theorem) rely on lemmas of such "extreme speciality."

---

[9] *Remarks on the Foundations of Mathematics*, G. H. von Wright, R. Rhees, and G. E. M. Anscombe, eds., Anscombe, trans. (Cambridge: MIT, 1978).

[10] *A Mathematician's Apology* (New York: Cambridge, 1992). Hardy was discussing recreational mathematics, but his points apply equally well to the combinatorial problem solved by Adleman.

Finally, it is often suggested that the problem with PDP is simply that it is probabilistic. Of course, this just begs the question. Whether or not there is any epistemic drawback to using probabilistic methods is precisely what is at issue.

II. CERTAINTY

It is often asserted, as it has been by Ivars Peterson,[11] that "absolute certainty is important because mathematics is constructed like a giant house of cards: one theorem piled on top of another...if one element were faulty, then the whole structure could come tumbling down" (*ibid.*, p. 15). As a result, it might be suggested that "our quest for absolute mathematical truths" requires methods that provide absolute certainty. Unfortunately, PDP clearly does not provide absolute certainty.

After using PDP, a mathematician is not absolutely certain that there is no HP in a given directed graph, since he is aware that there are two potential sources of error associated with PDP. First, it might happen that there are HPs in the graph, but that no strands of DNA representing these HPs are generated in step 1 (and, thus, no such strands remain after step 3). Second, it might happen that strands of DNA representing HPs are generated in step 1, but that they do not remain after step 3. In other words, in addition to the uncertainty that is built into (step 1 of) the algorithm, there is uncertainty due to the implementation of the algorithm. An HP generated in step 1 might not remain in the test tube after step 3, if the DNA technology does not perform as advertised. For instance, if the test tube is heated to too high a temperature in step 3, then all of the strands of DNA in the test tube could be destroyed.

Nevertheless, the problem with PDP cannot be that it does not provide absolute certainty. Despite the alleged importance of absolute certainty, methods acceptable to mathematicians do not provide absolute certainty either. In other words, as I shall argue below, this quality falls under category A.

The classification of all finite simple groups provides an example of a method acceptable to mathematicians which does not provide absolute certainty. In the 1980s, group theorists established that every finite simple group is either a member of one of 18 infinite families of finite simple groups or is one of 26 sporadic finite simple groups. But the mathematician Daniel Gorenstein[12] (the "architect" of the proof of the classification theorem) estimates that "the complete proof is scattered across the [approximately 15,000] pages of some 500 articles in technical journals" (*ibid.*, p. 104). As Peterson points out, because of the unprecedented size of this proof, "no one is sure that the [proof of the] classification theorem contains no errors whose consequences would be great enough to threaten the integrity and validity of the entire proof" (*op. cit.*, p. 280).

---

[11] *Islands of Truth* (New York: Freeman, 1990).

[12] "The Enormous Theorem," *Scientific American* (December 1985): 104-15.

DCP is also an acceptable method of establishing that there is no HP in a given directed graph.[13] This method does not provide absolute certainty either, however. It is possible for the computer to report mistakenly that there is no HP because of a software mistake or a hardware failure.

While the problem with PDP cannot be that it does not provide absolute certainty, it still might be suggested that it does not provide *enough* certainty. Much as I hate to admit it, PDP may not provide as much certainty as most of the methods acceptable to mathematicians. For instance, the reliability of PDP is undoubtedly somewhat lower than the reliability of DCP (Pentium processors notwithstanding).

Several mathematicians have raised serious doubts about the reliability of DNA computations when they are used to solve problems of any reasonable size.[14] For instance, it has been noted that determining whether or not there is an HP in a directed graph with 70 towns requires more DNA than can be reliably handled with current techniques. Nevertheless, there are reasons to believe that the problem with PDP (or with probabilistic methods in general) is not that it does not provide enough certainty.

First of all, if we did determine the lowest acceptable level of certainty, PDP (especially if used on fairly small graphs) might easily exceed this threshold. After all, many of the methods acceptable to mathematicians do not provide all that much certainty. For instance, according to Gorenstein, "perhaps the best evidence of [the validity of the proof of the classification theorem] is the fact that...no simple groups have been found since the announcement" (cited by Peterson, *op. cit.*, p. 280)!

Secondly, even if the reliability of PDP is somewhat lower than the reliability of DCP, this is only a contingent matter. After all, as Adleman points out, "today's electronic computers...are the product of a half century of extraordinary development. Molecular computers are less than a year old" (*ibid.*, p. 484).

---

[13] Computer proofs and 15,000 page proofs are certainly not the typical, nor the preferred, means of establishing mathematical truths. In fact, mathematicians are hard at work trying to simplify the proof of the classification theorem and the proof of the four-color theorem. Nevertheless, computer proofs and 15,000 page proofs are acceptable methods of establishing mathematical truths and it is legitimate for me to compare PDP with such methods. I am not arguing that PDP is as good as the beautiful and concise methods of proof for which mathematicians strive, but simply that PDP meets the minimal requirements for establishing mathematical truths.

[14] See "On the Potential of Molecular Computing," *Science*, CCLXVIII (1995): 481-84—a column with letters by M. and N. Linial; Y.-M. D. Lo, K. F. C. Yiu, and S. L. Wong; B. Bunow; Adleman (response).

Thirdly, the level of certainty provided is clearly not what leads mathematicians to reject probabilistic methods. According to the mathematician Carl Pomerance,[13] there is "a qualitative difference between probabilistic verification and mathematical proof that is important to mathematicians" (*ibid.*, p. 142). Even if the reliability of PDP is somewhat lower than the reliability of DCP, this does not necessarily represent a qualitative difference. In both cases, we have designed a physical device that can perform an algorithm and we have good inductive evidence that the physical device will perform as advertised.

Furthermore, the fact that mathematicians reject many extremely reliable probabilistic methods indicates that the level of certainty provided by PDP is not what is at issue (and that PDP would be unacceptable to mathematicians no matter how reliable it was). For instance, consider the probabilistic test developed by Michael Rabin[14] to establish that a number is prime.

It has been proven that, if a number *n* is composite (that is, if *n* is not prime), then over half of the natural numbers between 1 and *n* are witnesses to the compositeness of *n*. That is, if we perform a fairly simple numerical test on numbers between 1 and *n*, over half will testify to the compositeness of *n*.

Rabin's method exploits this fact. We pick at random a whole bunch of numbers between 1 and *n* and perform the numerical test on each of them. If *n* is composite, the overwhelming odds are that some of the numbers we test will testify to the compositeness of *n*. As a result, if none of the numbers we test testifies to the compositeness of *n*, the overwhelming odds are that *n* is prime.

Even though they do not consider it to be an acceptable method of establishing that a number is prime, mathematicians do think that Rabin's method is very reliable. In fact, it is even arguable that it is more reliable than any deterministic method of verifying that a number is prime.[17] In any case, in situations that only involve "wealth, health, and survival," Rabin's method is one of the preferred tests for primality (even among mathematicians).[18]

---

[13] "The Search for Prime Numbers," *Scientific American* (December 1982): 136-47.

[14] "Probabilistic Algorithm for Testing Primality," *Journal of Number Theory*, XII (1980): 128-38. I am not pulling a fast one by shifting the discussion from one probabilistic method (PDP) to another (Rabin's method) in the middle of my arguments. The details of PDP only come into play when I am arguing that PDP does not have particular epistemic drawbacks (that is, when I am arguing that qualities fall under category B). All of these arguments can easily be modified to show that Rabin's method does not have these epistemic drawbacks either (see, for example, footnote 29).

While the problem with PDP is not that it does not provide enough certainty, it still might be suggested that PDP does not provide the *right kind* of certainty. Since they do not provide absolute certainty, methods acceptable to mathematicians do not provide an unconditional guarantee that a mathematical claim is true, but they (unlike PDP) do provide a *conditional guarantee* that the claim is true. The methods acceptable to mathematicians arguably involve performing a procedure that is such that, if it is performed correctly, then the mathematical claim must be true.

For instance, DCP provides a conditional guarantee that there are no HPs in a given directed graph. As I noted above, software mistakes and hardware failures are a possibility. If no software mistakes or hardware failures occur, however, then it must be the case that there are no HPs in the graph. Also, writing down a derivation in first-order logic (starting from standard mathematical axioms) is a procedure that is such that, if it is performed correctly, then the mathematical claim must be true.

PDP, however, does not provide even a conditional guarantee. Even if the DNA technology performs as advertised, there still might be an HP in the graph. Specifically, we might have been unlucky and no strands of DNA representing the HPs in the graph were generated in step 1.

Nevertheless, there is reason to doubt that the problem with PDP is that it does not provide a conditional guarantee. Specifically, it is not clear that providing a conditional guarantee has much epistemic significance. For instance, consider the manner in which Kurt Gödel established the truth of the second incompleteness theorem.

In 1931, Gödel[19] published a paper presenting his two incompleteness theorems. The first states that for any consistent formal system powerful enough to formalize arithmetic, there are true but unprovable propositions. The second states that no consistent formal system powerful enough to formalize arithmetic can prove its own consistency. His 1931 paper contains a detailed proof of the first theorem, but only a sketch of a proof of the second theorem.

---

[17] See Pomerance, "Recent Developments in Primality Testing," *Mathematical Intelligencer*, III, 2 (1981): 97-105.

[18] See, for example, R. Rivest, A. Shamir, and Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," *Communications of the ACM*, XXI, 2 (1978): 120-26.

[19] "On Formally Undecidable Propositions of Principia Mathematica and Related Systems," *Collected Works, Volume I*, Solomon Feferman et alia, eds. (New York: Oxford, 1986), pp. 145-95.

Gödel's proof of the second theorem appeals to the fact that the proof of the first theorem is formalizable (in the formal system *P*). The only evidence that Gödel offered for the formalizability of the proof of the first theorem, however, was to say that the proof "used only the methods of definition and proof that are customary in classical mathematics, as they are formalized in system P" (*ibid.*, p. 193). Despite the fact that Gödel left out many of the details of a proof of the second incompleteness theorem, mathematicians felt that Gödel had established its truth.[30]

As I noted above, writing down a proof (or at least a first-order derivation) provides a conditional guarantee. But Gödel did not write down a proof. It is not even clear that he had worked through all the details in his head. Of course, Gödel did write down a sketch of a proof and, if he performed this procedure correctly, then the second incompleteness theorem must be true. Making wild guesses, however, provides a conditional guarantee in this same sense. If I guess correctly that the Goldbach conjecture is true, then the Goldbach conjecture must be true. Thus, there does not seem to be much epistemic significance to providing a conditional guarantee (at least in the sense that Gödel provided one).

Mathematicians clearly prefer methods that provide a conditional guarantee, but this preference may be due to aesthetic rather than epistemic considerations. In fact, as I shall argue below, even if we set aside problematic cases like the second incompleteness theorem and focus on cases like DCP, it is not clear that providing a conditional guarantee has much epistemic significance.

Using a method that does not provide a conditional guarantee does introduce an extra source of uncertainty. With deterministic algorithms, only the implementation of the algorithm is a source of uncertainty. With probabilistic methods, the implementation of the algorithm is a source of uncertainty *and* the algorithm itself is a source of uncertainty. Nevertheless, it is not clear that introducing an extra source of uncertainty is necessarily an epistemic drawback.

There are at least a couple of reasons why an extra source of uncertainty might be considered an epistemic drawback. First, it might be suggested that introducing an extra source of uncertainty means introducing more uncertainty. But there will often be less uncertainty with probabilistic methods, than with deterministic methods.[21] For one thing, probabilistic methods are typically much simpler than those deterministic methods which perform the same task. As a result, even though the (probabilistic) algorithm itself is a source of uncertainty, there is much less chance of an error in the implementation of the algorithm. Thus, despite an extra source of uncertainty, there will often be less uncertainty overall with probabilistic methods.

---

[30] See John W. Dawson, "The Reception of Gödel's Incompleteness Theorems," *Philosophy of Science Association 1984*, II (1985): 253-71. In fact, a detailed proof of the second theorem was not published until 1939.

Pomerance suggests a second way in which the extra source of uncertainty might be considered an epistemic drawback. He claims that the uncertainty involved in methods acceptable to mathematicians "can be reduced by studying the processes involved more deeply, by having other people check the proofs, by repeating calculations, etc. In fact there is no predetermined lower bound" (*ibid.*, p. 100). Pomerance comes to a different conclusion, however, with regard to probabilistic methods. For instance, if we use Rabin's method and test 100 potential witnesses to compositeness, then we face a probability of $1/2^{100}$ that there is a witness that we have not tested. Pomerance claims that this $1/2^{100}$ probability "just sits there. It is a positive constant, albeit small. It cannot be made smaller" (*ibid.*, p. 100).

Pomerance is just wrong, however. The probability of $1/2^{100}$ is not immutable; it can be reduced by simply testing more potential witnesses to compositeness.[22] Similarly, with PDP, the uncertainty due to the algorithm itself can be reduced by performing the algorithm more than once and (up to a point) by adding more copies of each town strand and each road strand in step 1. Both of these strategies will decrease the likelihood that we shall miss any HPs in the graph.

In conclusion, it is my contention that conditional guarantees fall under category C. In any case, if mathematicians want to insist that the extra source of uncertainty is the epistemically important qualitative difference between PDP and methods acceptable to mathematicians, we still need to know why it has epistemic significance and we need to know what to say about cases like the second incompleteness theorem.

---

[21] On Rabin's method, see, for example, Pomerance, "Recent Developments in Primality Testing."

[22] Reducing the uncertainty in this way requires doing more calculations instead of just checking over the calculations that we have already done. Nevertheless, there is no "predetermined lower bound" on how far we can reduce the uncertainty involved in probabilistic methods such as Rabin's method and PDP.

III. A PRIORI KNOWLEDGE

It is often claimed that mathematical knowledge is knowledge that can be justified independently of experience. In Adam Morton's[33] words, "we can learn mathematical facts just by thinking" (*ibid.*, p. 78) and without having to run experiments at a DNA laboratory. As a result, it might be suggested that the problem with PDP is that it does not give us a priori knowledge. In order to evaluate this suggestion, I shall consider three different definitions of a priori warrant (beginning with the most restrictive of the three).

Philip Kitcher's[34] definition of a priori warrant goes basically as follows: something is an a priori warrant if and only if it would justify our belief no matter what experiences we might have. Under Kitcher's definition, knowledge justified by an a priori warrant is incorrigible (that is, no possible further experiences could defeat an a priori justification). There are methods acceptable to mathematicians, however, which do not provide an a priori warrant under Kitcher's definition (as Kitcher himself points out in chapter 2). In other words, as I shall argue below, his definition of a priori warrant falls under category A.

Since mathematicians sometimes make mistakes, the warrant provided by checking a very long (purported) proof of a mathematical claim (such as the classification theorem) could be defeated by further experiences. For instance, we might find a counterexample to the claim (such as a new finite simple group). Also, even if the theorem is true, we might find out that there is an error in the proof. Further, even if there is no error in the proof, another mathematician (whom we take to be a more reliable proof checker than ourselves) might claim to have found an error in the proof. In all of these cases, we are no longer justified in believing that the claim is true.

The fact that not all methods acceptable to mathematicians are a priori under Kitcher's definition suggests that we need a weaker definition of a priori warrant such as the following: something is an a priori warrant if and only if it justifies our belief without appeal to any experiences we have had. In other words, an a priori justification does not require that we have had any particular experiences. Under this definition, knowledge justified by an a priori warrant is not necessarily incorrigible.

Some methods acceptable to mathematicians, however, do not provide an a priori warrant in this sense either. In fact, our justification for believing that the classification theorem is true still does not count as a priori under this definition, since it relies (even if you are Gorenstein) on the empirical fact that other mathematicians are reliable theorem provers. Also, the justification provided by DCP relies on the empirical fact that digital computers perform computations reliably. Even the justification provided by checking a relatively long proof by yourself relies on the empirical fact that blackboards, pieces of paper, and even your own memory are reliable devices for storing information.[15]

---

[33] *A Guide through the Theory of Knowledge* (Encino, CA: Dickenson, 1977). Of course, mathematical knowledge can also be justified with an appeal to experience. For instance, I might know that Fermat's Last Theorem is true only because I read about Andrew Wiles's proof in the *New York Times*.

[34] *The Nature of Mathematical Knowledge* (New York: Oxford, 1983), chapter 1.

Of course, it might be suggested that this is not the sort of empirical data which keeps a warrant from being a priori. Jody Azzouni[16] suggests one way in which we might distinguish between the type of empirical data to which an a priori warrant can appeal and the type to which it cannot. According to Azzouni, a warrant (which appeals to empirical data) might rely on empirical science or it might only rely on what he calls prescientific experience. With this distinction in mind, an Azzouniesque definition[17] goes as follows: something is an a priori warrant if and only if it justifies our belief without any appeal to empirical science. In other words, an a priori justification only relies on prescientific experience.

Roughly speaking, prescientific experience is what all of our common-sense knowledge about the world relies on. For instance, our confidence in our ability to judge that a pointer on a meter is at a particular location relies only on prescientific experience. The scientific theory that discusses the physiology of the human visual system helps to explain why we are able to judge the location of pointers reliably. Our confidence in our ability to judge the location of pointers reliably does not rely on this scientific theory. In fact, we could not even confirm this scientific theory if we were not already convinced of our ability to do things reliably such as judge the location of pointers.

Similarly, our confidence in the reliability of blackboards as recording devices relies on our everyday experience with blackboards. It does not, for instance, rely on our knowledge of the chemistry of chalk. Thus, if we have to record some intermediate steps on the blackboard while checking a long proof, then we still have an a priori warrant because we have only appealed to prescientific experience.

---

[15] On the proof of the four-color theorem, compare Michael Detlefsen and Mark Luker, "The Four-Color Theorem and Mathematical Proof," this JOURNAL, LXXVII, 12 (December 1980): 803-20.

[16] *Metaphysical Myths, Mathematical Practice* (New York: Cambridge, 1994), pp. 156-71.

[17] Azzouni's own definition of a priori warrant builds in the notion of surveying a derivation as well as the notion of prescientific experience. I shall consider the suggestion that an acceptable method must involve surveying a derivation in section IV below.

According to Azzouni, even his confidence in the reliability of floppy disks only relies on prescientific experience, since it "has nothing to do either with scientific doctrine (for I know virtually nothing about how floppies work) or with the authority of experts (for I do not know any experts in this area)" (*op. cit.*, p. 161). Similarly, a mathematician's confidence that digital computers perform computations reliably might rely only on prescientific experience. For instance, it need not rely on scientific knowledge about how transistors work any more than his confidence in the reliability of floppy disks does. As a result, DCP might rely only on prescientific experience and thus provide an a priori justification (under the Azzouniesque definition).[*]

Unfortunately, at the moment, PDP relies directly on empirical science. If a mathematician is convinced by PDP, it is because she accepts the scientific theory of how DNA operates. Of course, this is again only a contingent matter. If it is possible for an individual's confidence in the reliability of digital computers to rely only on prescientific experience, then eventually it should be possible for an individual's confidence in the reliability of DNA computations to rely only on prescientific experience as well.

In any case, the fact that mathematicians reject probabilistic methods that are a priori under the Azzouniesque definition indicates that this is not what is at issue. For instance, mathematicians do not consider Rabin's method to be an acceptable method of establishing mathematical truths. If DCP only relies on prescientific experience, then performing Rabin's method using a digital computer only relies on prescientific experience as well.

At this point, it might be suggested that we have gone too far. That is, in order to be able to count all methods acceptable to mathematicians as a priori, we have adopted a definition of a priori warrant which is much too weak. In fact, mathematicians might even be willing to throw out computer proofs if that is what it takes to keep them safe from probabilistic methods.

---

[*] Azzouni himself does not think that computer proofs, such as the proof of the four-color theorem, can rely only on prescientific experience since the calculations involved go beyond our ability to check prescientifically (*op. cit.*, p. 169-71). While we could not check all of the calculations, however, we can check prescientifically that the computer performs calculations correctly and that the program is correct. Of course, I have no great stake in outcome of this particular debate. If computer proofs turn out not to be a priori under the Azzouniesque definition, then the Azzouniesque definition simply falls under category A.

Nevertheless, I do not think that trying again to find a stronger definition of a priori warrant will help. Even if mathematicians were willing to throw out computer proofs, probabilistic proofs would still be a problem. For instance, suppose that I think up a whole bunch of possible paths through the graph and then verify that none of them is an HP. This is a probabilistic method of establishing that there are no HPs in the graph, but the justification that it provides is not dependent on experience. At least, it no more depends on experience than does checking (in an orderly fashion) every possible sequence of *N* towns in your head (which is a method acceptable to mathematicians).

In conclusion, if mathematicians want to insist that an a priori warrant is the epistemically important qualitative difference between PDP and methods acceptable to mathematicians, we shall need yet another definition of a priori warrant. It does not appear, however, that the notion of 'independence from experience' will cut a division between PDP and the methods acceptable to mathematicians.

IV. PROOF

Mathematicians typically establish that a mathematical claim is true by writing down a proof. In other words, they try to construct an object whose existence implies that the mathematical claim is true (or at least that it follows from standard mathematical axioms). It might be suggested that the problem with PDP is that it does not provide a proof. In this section, I shall discuss several refinements of this suggestion.

To begin with, it might be suggested (perhaps by someone like René Descartes) that the problem with PDP is that no human being actually surveys a proof that there is no HP in the graph. In other words, the suggestion is that in order to establish a mathematical truth, a mathematician has to bring each step of the proof before his mind and verify that it is a valid move.

There are methods acceptable to mathematicians, however, which do not involve a human being surveying a proof. In other words, this quality falls under category A. For instance, if we use DCP, no one will have surveyed a proof that there is no HP in the graph. Similarly, even though no human being has surveyed a proof of the four-color theorem, mathematicians seem to think that the truth of the four-color theorem has been established.

At this point, it might be suggested that the problem with PDP is that no proof is *constructed*. For instance, no one surveyed the proof of the four-color theorem, but the proof was constructed by the computer. There are methods acceptable to mathematicians, however, which do not involve the construction of a complete proof. As we saw above, Gödel did not write down a complete proof of the second incompleteness theorem. By the way, this is not an isolated example. According to Kitcher, "some theorems of analysis...never receive general proofs which are rigorous even by the standards of informal rigor which mathematicians accept" (*op. cit.*, p. 42). For instance, Stokes's theorem is applied in cases in which, Kitcher continues, "nobody doubts that the theorem is true...[even though]...no one has taken the trouble to prove it" (*op. cit.*, p. 42).

Of course, while Gödel did not fill in all of the details, he did write down a large part of the proof of the second incompleteness theorem. But PDP also involves the construction of a large part of a proof that there is no HP in a given directed graph. All that it takes to construct a complete proof is to verify that all paths of length *N* are generated in step 1. Thus, this quality falls under category B.[29]

In addition, PDP provides good evidence that the details of the proof can be filled in. As I argued above, PDP provides pretty good evidence that there is no HP in the graph. If there is no HP in the graph, then it is a trivial (but perhaps tedious) matter to construct a complete proof of this fact (for example, we could simply use DCP). Thus, this quality also falls under category B.[30]

It turns out that PDP and the manner in which Gödel established the truth of the second incompleteness theorem have a very similar structure. In both cases, part of a proof of a mathematical claim is constructed, but some of the details are left out. In both cases, however, good evidence is provided that these details could be filled in.

At this point, it might be suggested that the problem with PDP is that it does not provide the *right sort* of evidence that these details could be filled in. For instance, it has been suggested to me that Gödel's proof sketch not only provided good evidence that the details could be filled in, it also allowed mathematicians to "see that" the details could be filled in. In other words, the problem with PDP is supposed to be that it does not allow mathematicians to "see that" the details could be filled in.

This suggestion does identify a phenomenological distinction between PDP and the methods acceptable to mathematicians. That is, the experience that the mathematician has is different in the two cases. The methods acceptable to mathematicians do seem to provide mathematicians with a feeling of certainty which PDP does not. PDP allows a mathematician to imagine easily that she might be unable to fill in the details (since she might have been unlucky and no strands of DNA representing the HPs in the graph were generated in step 1).

---

[29] Since over half of the numbers between 1 and n are witnesses to the compositeness of n (if n is composite), showing that half of these numbers are not witnesses constitutes a proof of primality. Rabin's method involves the construction of a substantial part of such a proof.

[30] It is sometimes suggested that mathematicians have no use for probabilistic methods because they are interested in learning about deductive relationships between propositions and not probabilistic relationships. Since PDP provides evidence for the existence of a proof, however, it does give us information about a deductive relationship.

It is not clear, however, that this suggestion identifies an important *epistemic* distinction.[21] The phenomenology of an experience does not necessarily have any epistemic impact. For instance, writing down a derivation in first-order logic may make my scalp tingle, but this does not imply that my belief that there are no HPs in a graph is better justified (or that I shall have fewer flakes) than if I had just used PDP.

Of course, there are cases where the phenomenology of an experience does have epistemic impact. As Roderick M. Chisholm[22] notes, "if S feels sad, then S is at least as justified in believing that he feels sad as he is in having any other belief" (*ibid.*, p. 19). It is my contention, however, that "seeing that" does not have any epistemic impact and that it is only a phenomenological distinction between PDP and the methods acceptable to mathematicians.

At the moment, we do not have a precise characterization of how "seeing that" operates and when it can legitimately stand in for actually filling in the details of a proof. Even in the absence of a precise characterization of "seeing that," however, there are reasons to think that "seeing that" does not provide mathematicians with good grounds for rejecting PDP.

First of all, there are cases where using "seeing that" to justify the belief that the details of a proof can be filled in is not acceptable to mathematicians. According to G. Polya,[23] for instance, the young Sir Isaac Newton "read the theorems [of Euclid], saw that they were true, and omitted the proofs. He wondered why anybody should take pains to prove things so evident" (*ibid.*, p. 215). But mathematicians (including the older Newton) seem to consider this to be unacceptable mathematical practice. As a result, it is not clear why "seeing that" is such a crucial distinction between probabilistic methods and the methods acceptable to mathematicians.

---

[21] For instance, Gödel might simply have relied on the (probabilistic) evidence that he had formalized several other arguments which "used only the methods of definition and proof that are customary in classical mathematics" without any difficulty.

[22] *Theory of Knowledge* (Englewood Cliffs, NJ: Prentice-Hall, 1989, 3rd ed.).

[23] *How To Solve It* (Princeton: University Press, 1957, 2nd ed.).

Secondly, "seeing that" is not always a very reliable method of establishing that the details of a proof can be filled in. For instance, suppose that a mathematician wants to prove that a particular theorem follows from the axioms of Zermelo-Fraenkel with choice (ZFC). Also, suppose that her concept of set is not completely captured by the axioms of ZFC.[24] Finally, suppose that she only writes down a sketch of the proof on the grounds that she can "see that" the details could be filled in. In such a case, even if her mathematical intuition guides her unerringly to true facts about sets, it might easily happen that the theorem does not follow solely from the axioms of ZFC.[25]

Of course, someone might object that the discrepancy between ZFC and the mathematician's concept of set can lead to error when the mathematician only "sees that" the theorem is true, but not when she also "sees that" the details of the proof could be filled in. But consider Euler's[26] attempt to prove that "it is impossible to find any two cubes, whose sum, or difference, is a cube" (*ibid.*, p. 38).

In the course of his argument, Euler stated that "if a product of two factors, that are prime to each other, be a cube, each of itself must necessarily be a cube" (*ibid.*, p. 39) and he applied this statement to numbers of the form $a + b\sqrt{-3}$ where a and b are integers. Even though he did not offer a proof that this statement holds for numbers of this form, it is arguable that Euler (much like Gödel) could "see that" the details of a proof could be filled in. First, as the mathematician Harold M. Edwards[27] points out, Euler knew that numbers of the form $a + b\sqrt{-3}$ "form a number system much like that of the integers. In either case adding, subtracting or multiplying any two numbers in the system gives another number in the system, whereas dividing two numbers usually does not" (*ibid.*, p. 112). Second, he knew how to prove that "the product of two relatively prime positive integers can be an nth power only if each of the integers is an nth power" which can easily be "derived from the unique factorization of integers into primes" (*ibid.*, p. 112). Since numbers of the form $a + b\sqrt{-3}$ can sometimes be factored into primes in more than one way, however, Euler (unlike Gödel) would have been unable to fill in the details that he left out. Thus, "seeing that" *the details of a proof can be filled in* is not always a very reliable method of establishing that the details of a proof can be filled in.[illegible]

---

[24] The fact that Gödel thought that the continuum hypothesis (which is undecidable in ZFC) was false indicates that ZFC did not completely capture his concept of set (see "What is Cantor's Continuum Problem?" *Collected Works, Volume II*, Solomon Feferman et alia, eds. (New York: Oxford, 1990), pp. 184–85).

[25] Just such a mischance may have befallen Devlin in his attempt to establish that a particular result follows from an axiomatization of set theory weaker than ZFC (see Lee J. Stanley, Review of *Constructibility* by Keith J. Devlin, *Journal of Symbolic Logic*, LII, 3 (1987): 864-67, here p. 866).

[26] Quoted in D. J. Struik, *A Source Book in Mathematics, 1200-1800* (Cambridge: Harvard, 1969).

[27] "Fermat's Last Theorem," *Scientific American* (October 1978): 104-22.

In any case, as long as we do not have a precise characterization of "seeing that," there will be reason to doubt that mathematicians have good grounds for rejecting PDP. Compare what Charles Chihara[illegible] had to say about Gödel using mathematical intuition to justify his belief in the axioms of set theory:

> There is something suspicious about an argument for the existence of sets that rests upon data of so unspecified and vague a nature, where even the most elementary sorts of controls and tests have not been run. It is like appealing to experiences vaguely described as 'mystical experiences' to justify belief in the existence of God (*ibid.*, p. 215).

In conclusion, it is my contention that "seeing that" falls under category C. In any case, if mathematicians want to insist that "seeing that" is the epistemically important qualitative difference between PDP and methods acceptable to mathematicians, we still need to know why it has significant epistemic impact. At the very least, mathematicians should not be very comfortable as long the purported epistemic advantage of the methods acceptable to mathematicians over probabilistic methods "rests upon data of so unspecified and vague a nature."

Before leaving the topic of proofs behind, I want to consider one final suggestion. Above, I argued that PDP provides good evidence that a proof can be constructed. Now, PDP provides good evidence that there is a proof of a mathematical claim because it provides good evidence that the claim is true. But the methods acceptable to mathematicians typically work the other way around. That is, the method provides good evidence that a mathematical claim is true because it provides good evidence that there is a proof of the claim. As a result, it might be suggested that the problem with PDP is that it gets the epistemic order backward. As I shall argue below, however, mathematicians reject probabilistic methods that do not get the epistemic order backward. This fact indicates that epistemic order is not what is at issue.

[illegible] Of course, someone might suggest that Euler did not "see that" the details could be filled in *in the right sort of way*. In that case, however, since the manner in which Euler established his result was acceptable to mathematicians (at least until the gap was discovered), "seeing that" simply falls under category A.

[illegible] "A Gödelian Thesis regarding Mathematical Objects: Do They Exist? And Can We Perceive Them?" *Philosophical Review*, xci, 2 (1982): 211-27.

Computer scientists have recently developed an algorithm that can rapidly verify (with a small probability of error) the validity of any (suitably formalized) mathematical proof.[40] Now, there are acceptable uses to which this *probabilistic proof checker* (PPC) can be put. For instance, a PPC could legitimately be used to confirm that the proof of the four-color theorem which has been constructed is indeed correct. Similarly, even though the fact that "no simple groups have been found since the announcement [of the proof of the classification theorem]" does not establish the truth of the classification theorem, it is good corroborating evidence. But below I shall use a PPC to construct a probabilistic method which mathematicians will reject but which will not get the epistemic order backward.

Suppose that we write a program that generates (in an orderly fashion) all possible finite sequences of finite strings in a given alphabet. Next, suppose that we feed the output from this program into a PPC. If the PPC reports that one of these sequences is a proof, then we have good (probabilistic) evidence that a particular mathematical claim is true. Furthermore, this method provides good evidence that the claim is true precisely because it provides good evidence that there is a proof of the claim. Since this method provides *only probabilistic* evidence that there is a proof, however, mathematicians will not (I would venture to guess) consider this to be an acceptable method of establishing mathematical truths.

V. CONCLUSION

Much as I hate to admit it, mathematicians may have good grounds for rejecting PDP. At the moment, PDP may not provide enough certainty (see section II above) or may rely too heavily on empirical science (see section III above). These are arguably only contingent matters, however; we have not found an epistemically important *qualitative* difference between PDP and the methods acceptable to mathematicians. As a result, there is reason to doubt that mathematicians have good grounds for an across-the-board rejection of probabilistic methods as a means of establishing mathematical truths.[41]

---

[40] See Laszlo Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy, "Checking Computations in Polylogarithmic Time," *Proceedings of the 23rd ACM Symposium on Theory of Computing* (1991): 21-31. Again, for the same reason that I gave in footnote 16, I am not pulling a fast one by shifting the discussion to yet another probabilistic method.

[41] If they do not have good grounds (and they want to determine what is acceptable mathematical practice solely on the basis of epistemic considerations, compare the opening section above), then there are at least two modifications that would make mathematical practice consistent. Mathematicians could start using some probabilistic methods *or* stop using some of the methods that are currently in use.

FAINTHEARTED CONDITIONALS*

Heat a dry piece of wood thoroughly and you can expect it to burn. Why expect this but not, say, that it will dissolve when you put it in water? Wood is combustible but not soluble in water. That it is combustible might be thought to guarantee the truth of the conditional sentence:

> If it is heated thoroughly, then the wood will burn.

If it did, an obvious suggestion could be made: that the wood will burn would follow with logic; that it will dissolve would not follow. Reasonable expectations, we could say, are those entailed by the dispositional claims we make.

But the fact that wood is combustible does not guarantee truth of the above conditional: where, for example, there is insufficient oxygen present, this conditional is false, even though the wood remains combustible. The same disposition does guarantee the truth of the weaker conditional:

> If it is heated thoroughly, then provided conditions are suitable the wood will burn.

This sentence is true whether or not there is enough oxygen present, since conditions suitable for the combustion of wood include the presence of sufficient oxygen in the surrounding atmosphere. But this weaker conditional does not entail that the wood will burn, given only that it has been thoroughly heated. It seems that logic does not separate this reasonable expectation from the unreasonable expectation that the wood will dissolve in water, which does not follow either.

If-then sentences qualified by 'provided conditions are suitable' belong together with those qualified by 'under normal circumstances' or by 'other things being equal'. These qualifications suggest that exceptions are open ended and cannot exhaustively be listed. Thus the piece of wood will not catch fire if there is insufficient oxygen or if it starts to rain, but also under a very mixed bag of other circumstances. Call if-then sentences with any of these provisos *fainthearted conditionals*. Without such a qualification call them *boldhearted conditionals*.

---

* I am grateful to Michael Devitt, Jeff Horty, Sam Kerstein, Georges Rey, and Michael Slote for their comments on earlier versions of this paper.

0022-362X/97/9404/187–211

These provisos differ substantially from each other. For example, dispositional facts seem to support the truth of conditionals qualified by 'provided conditions are suitable' but not of those qualified by 'other things being equal'. Imagine a piece of wood under anoxic conditions—in a charcoal furnace, for example. You can heat it all you like but it will not burn, other things being equal. Other things being equal, the furnace will function properly, the oxygen level will stay below that required for combustion, and, instead of burning, the wood will turn into charcoal. In this setting, the wood remains combustible, but

> If it is heated thoroughly, then other things being equal the wood will burn.

is false.

Such differences notwithstanding, from a formal perspective fainthearted conditionals have much in common. In particular, they all provide a basis for reasonable expectations. I shall suggest a solution to *the problem of fainthearted conditionals*: the problem of distinguishing reasonable expectations, which have a basis in the fainthearted conditionals we accept, from unreasonable expectations. I shall illustrate this solution with a range of examples.

'Reasonable', in the sense I mean it here, is a relative term. An expectation can be reasonable in relation to one body of beliefs, but unreasonable in relation to another. This sense must be distinguished from another in which an expectation can be *simply* reasonable or unreasonable, independently of one's beliefs. For example, someone who believes that grass clippings cling to magnets can, in my sense, reasonably expect clippings to cling after touching them with a magnet. This expectation has a basis in his belief. But provided there is nothing funny about the clippings or the magnet—they are not covered with glue or anything like that—it is simply unreasonable. This expectation has a basis in an eccentric belief, but not in anything there is good reason to believe.

The problem of fainthearted conditionals has ramifications in explanation and prediction in science. To take just one example, empiricists think scientific theories include what Carl Hempel calls *interpretative statements*, linking the theoretical vocabularies of theories to empirical phenomena. Operational definitions and Rudolf Carnap's reduction sentences are examples. Now, some interpretative statements are fainthearted conditionals, and telling just which phenomena they help explain and predict is the problem of fainthearted conditionals.

An example of Hempel's shows why. According to Hempel,[1] the theory of magnetism has an interpretative principle like:

> If a bar is a magnet, then iron filings will cling to it.

Reasoning abductively from the observation that iron filings are clinging to a given bar, this principle supposedly enables us to infer the best explanation: the bar is a magnet. That, Hempel thinks, is how we identify it as such. Underlying this abductive reasoning is an inference in the other direction, leading from the hypothesis that the bar is a magnet to the expectation that the iron filings will cling. Now, this explanatory inference cannot be valid, since magnetic theory cannot *entail* that iron filings will cling to a given magnet. They will cling only if they are not at the same time in the grip of an immense competing gravitational field, they are not glued to the table, and so on, and it is not up to magnetic theory to decide all of that. The interpretative principle I stated does entail that the iron filings will cling, though, so it cannot be a part of magnetic theory.

But if magnetic theory does not contain this principle, then which does it contain? A slightly weaker one, specifying that there is no such gravitational field and no glue? Thus diminished, the principle will not entail that iron filings will cling to a bar, given only that it is a magnet. Still it will not do. Weakening the principle in this way is the first step in retreating altogether from the idea that scientific theories have interpretative principles, since it is easy to imagine other interfering factors, and yet others, which will require further weakening, and further weakening still. It is impossible to list all circumstances that can interfere, so if interpretative principles must list them, then there just *are* no interpretative principles.

Instead of weakening the bold-hearted principle, replace it with a fainthearted alternative like:

> If a bar is a magnet, then provided conditions are suitable iron filings will cling to it.

This principle will not entail that iron filings will cling to any given magnet. There will be no need to follow the interpretative principle through a succession of ever more feeble approximations. But with this gap stopped, another opens up: if interpretative principles are fainthearted, it is not clear how we can use magnetic theory to predict and explain such things as iron filings clinging to magnets, in ordinary cases where there is no evidence of immense gravitational fields, glue, or whatever. This is where a solution to the problem of fainthearted conditionals will come in.

---

[1] "Provisos: A Problem Concerning the Inferential Function of Scientific Theories," *Erkenntnis*, XXVIII (March 1988): 147-64.

The consequence of this interpretative principle for any given bar will ordinarily be put in the indicative mood, evoking a doxastic context in which filings might cling:

> If *this* bar is a magnet, then provided conditions are suitable iron filings will cling to it.

In a different setting, where for whatever reason they clearly will not cling, we might say instead:

> If this bar is a magnet, then were conditions suitable iron filings would cling to it.

The subjunctive mood of proviso and consequent reflects their status as doxastic *impossibilities*; in such a setting, of course, we cannot reasonably expect filings to cling to the bar.

Perhaps surprisingly, there are other conditionals with doxastically impossible provisos and consequents that do provide a basis for reasonable expectations. There are, for example, what Geoffrey Joseph[3] called *ceteris absentibus conditionals*. Joseph considered this one as a rendition of Coulomb's law:

> If p is a point charge, then were all other factors absent, p would be surrounded by a spherically symmetrical electromagnetic field.

The interesting thing is that Coulomb's law provides a basis for predictions about the electromagnetic field around a given charge p, even though we know that other factors must be present and that the field around p cannot be spherical. We can reasonably expect that the field is as *nearly* spherical as the interfering factors are negligible.

The solution to the problem of fainthearted conditionals I shall suggest has its origin in the following idea: an expectation is reasonable if, as far as we can tell, we can put aside the possibility that it is false. When can we put aside a possibility, as far as we can tell? That is the main question I have to answer and I shall not get to it properly until sections II and III. Meanwhile, an example will illustrate the answer I have in mind.

---

[3] "The Many Sciences and the One World," this JOURNAL, LXXVII, 12 (December 1980): 773-91.

Sitting at a fire, someone throws in a dry piece of wood. To justify the expectation that it will burn, we can reason as follows: of course, it is *possible* that the wood will not burn. But wood is combustible, and where a combustible thing does not burn after being thoroughly heated, there is some irregularity. Now, in addition to the possibility that the wood will not burn, as far as we can tell there is also the possibility that it *will* burn. Of the two possibilities, the first is *less regular*: it involves an irregularity that the second possibility does not involve. I shall say that the possibilities we can put aside, as far as we can tell, are ones like this: among our doxastic possibilities, they are less regular than some.

To characterize the logical consequences of fainthearted conditionals, or of any other sentences for that matter, we must first interpret them. The same holds for reasonable expectations: to draw a line between the expectations which do and those which do not have a basis in fainthearted conditionals, we must first have an understanding of what it takes for something to hold, provided conditions are suitable (or under normal circumstances, or other things being equal). In section II, I shall explicate the notion of comparative regularity, with which I shall draw this line, in terms of fainthearted conditionals. I shall say that there is an irregularity if something does not hold which would have held, had conditions been suitable (or circumstances normal, or other things equal). So my account of comparative regularity depends on an interpretation of fainthearted conditionals.

Notice the difference between comparative regularity and two other notions with which it is naturally contrasted: Peter Gärdenfors's[3] comparative epistemic entrenchment, which via the Ramsey test and the Alchourrón-Gärdenfors-Makinson theory of belief revision can be used to give an account of the assertibility conditions of indicative conditionals, and David Lewis's[4] notion of comparative similarity, which underlies his semantics for subjunctive conditionals. Whereas these are primitive notions on which interpretations of conditionals depend, comparative regularity is a derivative notion which depends on an interpretation of conditionals.

In the coming section, I shall suggest an interpretation of fainthearted conditionals. This will serve two purposes. First, the interpretation can be grasped independently of the notion of comparative regularity I shall rest on it. Standing alone, but supporting compara-

[3] *Knowledge in Flux* (Cambridge: MIT, 1988).

[4] *Counterfactuals* (Cambridge: Harvard, 1973).

tive regularity and the rest of my account of reasonable expectations, it provides them with a basis in things that can be understood independently. Second, having an interpretation of conditionals at hand makes the discussion more concrete, enabling examples to be analyzed with rigor.

Unfortunately, it also makes things more technical. You can pass over section I if, when it comes to the examples, you will take at face value my claims about the possibilities left open by bodies of beliefs. You can also pass over this section if you prefer an interpretation other than mine. For while my account of reasonable expectations presupposes some or other interpretation of fainthearted conditionals, it does not depend essentially on the one I shall suggest. Another interpretation will do just as well, provided it can be used to order possibilities according to their degree of regularity.

I. AN INTERPRETATION OF FAINTHEARTED CONDITIONALS

The idea underlying the interpretation I shall suggest is that the conditional:

> If it is heated thoroughly, then provided conditions are suitable the wood will burn.

is correct only if there are conditions suitable for the heated wood to burn. Call them collectively the *implicit condition*. Furthermore, what one says by uttering this sentence can also be put:

> If it is heated thoroughly and the implicit condition holds, then the wood will burn.

Before elaborating on this idea, I must deal with an apparent problem. Is it not trivial that there is a condition suitable for the wood to burn? It is the condition that obtains, well, whenever the wood will, in fact, burn. Since it is tautologous that the wood will burn if it is thoroughly heated and this condition holds, is what one says by uttering the above sentence not merely fainthearted, but completely vacuous?

I think not, but to see why not it is necessary to say more about implicit conditions. I suggest that what saves fainthearted conditional utterances from vacuity is that they are not merely supposed to convey truths. In saying the earlier sentence, one means to express a useful little law: something which is not only true, but also useful in predicting and explaining the outcome of heating the wood. For this reason, not all conditions count which will be followed by the wood's burning. Additionally, the implicit condition must avail us in predicting the burning before the event, and in explaining it afterward.

The trivial condition that the wood will burn clearly does not fill the bill. For one thing, unless it can be characterized in some more interesting way, we have no way of knowing whether it holds, apart from waiting to see whether the wood burns, and by the time that happens it will be too late for a prediction. Another condition seems to lend itself to predictions and explanations, however: the presence of oxygen in the surrounding atmosphere and its not being too wet. Plausibly, what one says, when one uses the above sentence, is that, if the wood is heated thoroughly, there is sufficient oxygen, and it is not too wet, then the wood will burn.

For an example of a conditional that is not correct because there is no such useful condition, suppose someone holds up a coin and says:

> If I toss this coin, then provided conditions are suitable it will land heads.

Now, I do not know whether coin tossing is truly indeterministic. If it is, then certainly there is no condition that will be useful in predicting the outcome. But even if coin tossing is deterministic, then still the determining factors are, in an everyday context, well beyond our powers of control and observation. Outcomes are for practical purposes, unpredictable (which is why we toss coins). In this case, too, there is no implicit condition that will be useful in predicting and explaining the outcome of the toss. The conditional is therefore not correct.

Putting aside the question of their apparent vacuity, I shall return now to the interpretation of fainthearted conditionals. I shall proceed as follows. First, I shall present an interpretation of boldhearted conditionals. Then I shall extend it to fainthearted conditionals by introducing implicit conditions.

There are two main paradigms in the semantics of conditionals. One is the epistemological analysis suggested originally by Frank Ramsey and elaborated by Isaac Levi,[3] which is concerned with the assertibility conditions of conditionals. I once defended this analysis against an argument due to Gärdenfors,[4] which purported to show that it cannot account for subjunctives in a realistic way, and if subjunctives like Joseph's ceteris absentibus conditionals do indeed have epistemological interpretations, then this semantic paradigm can serve as the basis for a solution to the problem of fainthearted conditionals. Here, I have chosen to work within the other main paradigm, however, which is concerned with the truth conditions of conditional sentences. Apart from anything else, the notion that conditionals bear truth values has the advantage, for my purpose, of its familiarity.

---

[3] Ramsey, *The Foundations of Mathematics* (New York: Routledge, 1931); and Levi, "Subjunctives, Dispositions and Chances," *Synthese*, xxxiv (1977): 423-55.

[4] See my "Epistemic Semantics for Counterfactuals," *Journal of Philosophical Logic*, xxi, 1 (1992): 33-62. See also Gärdenfors, "Belief Revisions and the Ramsey Test for Conditionals," *Philosophical Review*, xcv (January 1986): 81-93.

Some supporters of a truth-conditional interpretation think the indicative is just the material conditional. If they are right, then there are problems with inferences—the so-called *paradoxes of implication*—which are valid, although as arguments they will convince no one. Thus, while $\neg A$ entails the material conditional $A \rightarrow C$ for any sentence C at all, no one will accept that the cows are on their way home if the barn door is open, on being informed, merely, that the barn door is closed. Paul Grice[7] argued that such arguments violate general norms of conversation and are unacceptable on pragmatic grounds.

Many others[8] say the indicative conditional is not truth functional but is on a par with the subjunctive. Robert Stalnaker suggested this idea and developed it technically with Richmond Thomason;[9] this account of indicatives is closely related to Lewis's treatment of subjunctives (*op. cit.*). Both deal only with bold-hearted conditionals, however. I shall borrow freely from them, introducing implicit conditions as missing antecedents. In this way, I shall provide fainthearted conditionals with truth conditions in possible worlds.

Truth conditions in possible *whats*? Worlds: ones other than the actual world. Worlds where things are as things might actually be, or have been. I shall quantify over possible worlds. I shall define truth conditions relative to them. I shall speak as if they exist. I shall do these things without further apology because it does not matter much if there are none. State descriptions will do just as well.[10] So will complete consistent novels,[11] ways-things-may-be-or-could-have-been,[12] and other things that have been suggested in their stead.[13]

---

[7] In his 1967 William James Lectures; see *Studies in the Way of Words* (Cambridge: Harvard, 1989).

[8] For example, Steven Read recently presented an argument to this effect in "Conditionals Are Not Truth-functional: An Argument from Peirce," *Analysis*, LII, 1 (January 1992): 5-12.

[9] Stalnaker, "A Theory of Conditionals," in *Studies in Logical Theory, American Philosophical Quarterly*, Monograph 2 (Cambridge: Blackwell, 1968), pp. 98-112; Stalnaker and Thomason, "A Semantic Analysis of Conditional Logic," *Theoria*, XXXVI (1970): 23-42; and Stalnaker, "Indicative Conditionals," *Philosophia*, V (July 1975): 269-86.

[10] Suggested by Rudolf Carnap in *Meaning and Necessity* (Chicago: University Press, 1947).

[11] Suggested by Richard Jeffrey in *The Logic of Decision* (Chicago: University Press, 1983, 2nd ed.)

The idea shared by Lewis's and Stalnaker's truth conditions is this: a bold-hearted conditional holds true, in any given possible world of evaluation, if the consequent holds in those antecedent worlds which are most like the world of evaluation.[14] The idea underlying my adaptation is this: a fainthearted conditional holds true if the consequent holds true in those antecedent worlds which, among worlds where the implicit condition holds, are most like the world of evaluation.

To see what likeness among worlds has to do with things, consider a dry stick of wood. It is combustible and:

> If it is heated thoroughly, then provided conditions are suitable this stick will burn.

is true. Now, when we speak of conditions being suitable for the combustion of wood, we have in mind environmental conditions like the presence of oxygen, not the properties of this particular stick. For this reason, the above sentence would not be true if in evaluation one had to look at *all* possible worlds where the stick is heated under suitable conditions, no matter how unlike the actual world they are. In some such worlds, the stick is waterlogged, say, though actually it is not. In these worlds, it will not burn when heated. Instead, it will sizzle and give off steam. In evaluating this sentence, only worlds more like the actual world ought to matter, where the stick is as it actually is: dry.

Lewis introduced *comparative-similarity* relations to capture the idea that worlds can differ in the extent to which they resemble any given world. Let '$u \leq_w v$' stand for "the world u is at least as similar to w as the world v is" (*op. cit.*, p. 48).[15] Where $u \leq_w v$ but not $v \leq_w u$, say 'u is more similar to w than v is', and write '$u <_w v$'. Using Stalnaker's corner '>' as the bold-hearted conditional connective, the following truth conditions embody the idea that $A>C$ is true if C is true at the closest *A*-worlds.[16]

---

[12] Suggested by Stalnaker in "Indicative Conditionals," *Philosophia*, v (July 1975): 269-86.

[13] They will do just as well for my present purpose, which is to describe the form of the truth conditions of fainthearted conditionals. In *On the Plurality of Worlds* (Cambridge: Blackwell, 1986), Lewis argued that such "ersatz" possible worlds will not do as well for other purposes.

[14] Kit Fine and Michael Slote have suggested counterexamples to an analysis involving the comparative similarity of possible worlds. Considering sentences like:

> If Nixon had pressed the button, there would have been a nuclear holocaust.

They argue that worlds differ more from actuality where there was a holocaust, after Nixon pressed the button, than do worlds where through some slight divergence from actuality—the button was disconnected, say—there was no holocaust and life went on very much as it actually did. See Fine's review of Lewis's *Counterfactuals*, *Mind*, LXXXIV (1975): 451-58. See Slote, "Time in Counterfactuals," *Philosophical Review*, LXXXVII, 1 (January 1978): 3-27. See also Lewis's discussion in "Counterfactual Dependence and Time's Arrow," *Noûs*, 13 (November 1979): 455-76.

> $A>C$ is true at world w just in case for any *A*-world v, there is an *A*-world u such that (1) $u \leq_w v$, and (2) for any *A*-world t for which $t \leq_w u$, C holds true at t.

Thus, $A>C$ is true at w if, starting at any given *A*-world and moving closer to w, there comes a point after which each *A*-world is a *C*-world.

An argument is *valid* if the premises guarantee the conclusion. In terms of possible worlds, a valid argument is one with no *counterexample*. There is no possible world where the premises are true, but the conclusion false. Now, which arguments are valid depends in part on the properties of $\leq_w$. One basic requirement is that no world is more similar to w than w itself; there is no v such that $v <_w w$. It is not difficult to see that this requirement validates modus ponens: any argument from premises A and $A>C$ to the conclusion C is valid. Equivalently, the axiom scheme $(A>C)\rightarrow(A\rightarrow C)$ is valid in the sense that every sentence of this form is true at every world. Constrain $\leq_w$ as Lewis did and, apart from his "vacuous" case, > has the same truth conditions as his counterfactual conditional (*op. cit.*, p. 48).

So far, we have an interpretation of bold-hearted conditionals. To extend it to fainthearted conditionals, it remains to introduce implicit conditions. The main question to answer is this: when we evaluate a fainthearted conditional, what determines the condition implicit in the proviso? To answer this question, it is important to take into account the fact that provisos, as usually expressed, are elliptical. Something is missing when we simply say 'provided conditions are suitable', since conditions are not simply suitable. They are suitable for the combustion of wood, or going sailing, or what have you. Restoring the modifier, we have:

> If it is heated thoroughly, then provided conditions are suitable for the combustion of wood, the wood will burn.

---

[15] You might expect to read '$u \geq_w v$' instead, but I shall use Lewis's notation to avoid confusing anyone who is used to it.

[16] An *A-world* is a world where A is true.

The same holds for other provisos. If this engine is warm, then under normal circumstances it will start easily. But circumstances are not simply normal. Restoring the modifier, we have:

> If this engine is warm and circumstances are normal with regard to the ease with which it will start, then it will start easily.

Let an *end* be anything for which conditions can be suitable. Let a *regard* be anything with regard to which circumstances can be normal. Sometimes conditions are suitable for one end, but not for another; sometimes circumstances are normal in one regard, but not in another. Evidently, what is implicit in 'provided conditions are suitable' and 'under normal circumstances' depends on the ends for which conditions are to be suitable and the regards in which things are to be normal.

Introducing ends and regards as elliptical elements raises a new question: How can we know which end or regard to restore to a proviso, when evaluating a conditional in which none is explicitly present? I shall not answer this question here. Anything but a cursory discussion would lead us far astray; besides, we do not really need an answer for our present purpose. In the coming examples, as in the previous ones, it will be clear enough which ends and regards modify provisos.[17]

The implicit condition of a fainthearted conditional also depends on the context in which it is evaluated. In a game of poker, it can happen that

> If I draw a card, then provided conditions are suitable I shall make a straight.

Now, which conditions are suitable for making a straight depends on my hand. With, say, a five, a jack, a queen, a king, and an ace, the implicit condition is that the card I am to be dealt is a ten. With a different hand, there is a different implicit condition.

---

[17] Examples like these suggest that the end to be restored to 'provided conditions are suitable' depends simply on the conditional consequent:

> If this piece of wood is heated thoroughly, then provided conditions are suitable it will burn.
>
> If this piece of wood is heated thoroughly, then provided conditions are suitable it will turn into charcoal.

In the first sentence, conditions are to be suitable for the combustion of wood; in the second, for making charcoal. In both, conditions are to be suitable, roughly, for truth of the consequent. All later examples follow this pattern.

Things are not in general this simple, however. Suppose that

> If you want to go sailing, then provided conditions are suitable that is what we shall do.

Suppose, furthermore, that if we are going then one of us will have to bring provisions. Then when I promise you that:

> If you want to go sailing, then provided conditions are suitable I shall bring along a sack lunch.

it is clear that I am not hedging in case conditions are not suitable for bringing along a sack lunch. Conditions will almost certainly be suitable for *that*. I am hedging in case conditions are not suitable for sailing.

Finally, the implicit condition depends on the explicit antecedent of the conditional. The sentence:

> If I draw two cards, then provided conditions are suitable I shall make a straight.

differs from the previous example only in its antecedent. But its implicit condition concerns two cards that I am to be dealt, not one.

Now, a *proviso function* $\wp$ will model the role of these three determiners of the conditions implicit in 'provided conditions are suitable'. Analogous functions will do the same for other provisos. In accordance with the discussion, $\wp$ maps worlds (which are maximally inclusive contexts), propositions (which are sets of worlds), and ends (which are abstract objects) onto propositions. Intuitively speaking, $\wp(w,p,e)$ is the proposition implicit in the proviso of a conditional used in the context w, whose antecedent expresses p, and which concerns suitability for e. Where A expresses p, write '$\wp(w,A,e)$' instead of '$\wp(w,p,e)$'.[18] For example, let $@$ be the actual world. Then:

> $\wp$ ($@$, 'this piece of wood is heated thoroughly', the combustion of wood)

stands for the proposition, implicit in the proviso of

> If this piece of wood is heated thoroughly, then provided conditions are suitable it will burn.

that conditions are suitable for the combustion of wood. It is the proposition that there is enough oxygen, it is not raining, and so on.[19]

There seem to be conditions suitable for the combustion of wood, or for going sailing, or for making a straight in poker. But, as I said earlier, in an ordinary context there are no conditions suitable for a tossed coin to land heads. For this reason, it would be realistic to allow

---

[18] The sentence A *expresses* proposition p if p is the set of worlds where A is true.

[19] If it is heated thoroughly, then provided conditions are suitable this piece of wood will burn. Also, if it is heated thoroughly, then provided conditions are suitable it will turn into charcoal. It is implicit in the proviso of the first conditional that there is enough oxygen for the combustion of wood. It is implicit in the proviso of the second conditional that there is not enough oxygen for this, so these two examples show that the implicit condition can depend on the conditional consequent. What does not follow is that the proviso function ought to include conditional consequents among its arguments. Instead, as I see it, the consequent plays a role in determining the elliptical end for which conditions are to be suitable. In the first sentence, they are to be suitable for the combustion of wood; in the second, for making charcoal. Conditional consequents can determine implicit conditions indirectly, through the mediation of these ends.

$\wp$ to go undefined for some worlds, propositions, and ends, but I shall not allow this. Letting $\wp$ be a partial function complicates things in ways that are quite well known, and have been explained elsewhere.[30]

I shall write '...$\rhd_e$...' to abbreviate:

> If..., then provided conditions are suitable for e,...

Later the triangle will stand for other fainthearted connectives, too. Call '$P(w,A,e)$-worlds' those worlds where A holds along with $\wp(w,A,e)$. The following truth conditions let $A \rhd_e C$ be true, at any given world w of evaluation, if C holds in the closest $P(w,A,e)$-worlds.

> $A \rhd_e C$ is true at world w just in case for any $P(w,A,e)$-world v, there is a $P(w,A,e)$-world u such that (1) $u \leq_w v$, and (2) for any $P(w,A,e)$-world t for which $t \leq_w u$, C holds true at t.

Thus, $A \rhd_e C$ is true at w if, starting at any given A-world where conditions are suitable, and moving closer to w, there comes a point after which each A-world where conditions are suitable is a C-world.

To illustrate the truth conditions of $A \rhd_e C$, and to prepare for later examples, let a *frame* be a set of possible worlds, along with a suitable comparative-similarity relation and proviso function. One class of frames that will repeatedly appear later, when we come to the examples, contains just two worlds, u and v. The proviso function $\wp$ is such that for each end e, $\wp(u,\{u\},e) = \wp(v,\{u\},e) = \{u\}$, while for any proposition p other than $\{u\}$, $\wp(u,p,e) = \wp(v,p,e) = p \cap \{v\}$. Let an arrow drawn from p to q mean that for each e, $\wp(u,p,e) = \wp(v,p,e) = q$. The frames in this class have the following structure in common:

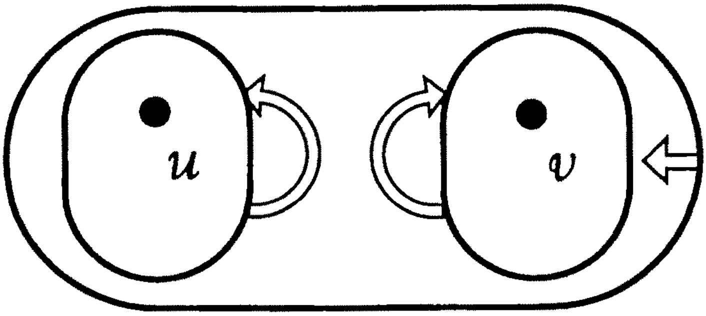


Figure 1

---

[30] There is a thorough technical treatment in Steven Blamey's chapter, "Partial Logic," in D. Gabbay and F. Guenthner, eds., *The Handbook of Philosophical Logic, Volume III* (Dordrecht: Reidel, 1986).

Fainthearted conditionals can be interpreted in the worlds of these frames by assigning truth values to their antecedents and consequents there. Notice that because in any such frame the values of are singletons (or sets containing single worlds), the truth values of fainthearted conditionals in worlds of these frames will be independent of the comparative-similarity relation. This follows immediately from the truth conditions for $\rhd$, since if $\wp(w,A,e)$ is a singleton, then $A\rhd_e C$ is true at w just in case the material conditional $A \rightarrow C$ is true at the world in $\wp(w,A,e)$.

Letting A be true in both u and v, for example, but letting C be true only in v, $A\rhd_e C$, A and $\neg C$ are true in u. This illustrates the main formal difference between faint- and bold-hearted conditionals: modus ponens does not hold for $\rhd$. The scheme $(A\rhd_e C) \rightarrow (A \rightarrow C)$ is not valid.

II. ALLOWED ARGUMENTS

Earlier, I introduced informally the idea that some doxastic possibilities are less regular than others. These less regular ones I shall call *unnecessarily irregular possibilities*. In this section, I shall tie them into the problem of fainthearted conditionals, using the following critical policy on counterexamples to draw the line between reasonable and unreasonable expectations: ignore any possibility that is unnecessarily irregular. This policy will produce expectations that logic alone does not produce, since even invalid arguments have a certain force if their counterexamples can be ignored. Such arguments have conclusions that follow on assuming the world is *not* unnecessarily irregular. Their conclusions, I shall suggest, are expectations with a basis in one's beliefs about the dispositional properties of things.

To make these ideas precise, I shall turn first to the question of what it is for a counterexample to be irregular, and unnecessarily so.

Things do not always go as they ought. In the actual world, a given dry piece of wood is combustible, but in spite of being thoroughly heated, it does not burn. In another world, possible but not actual, a piece of wood is soluble, not combustible. Still, though immersed, it will not dissolve. In another world, iron filings will under suitable conditions cling to a magnet, but instead they drop straight off all of them. In yet another world, it is not iron filings that tend to cling, but grass clippings. Even so, none of them does. These worlds are *irregular*. They are so in their own terms: they are worlds where things do not behave as they tend to *there*.

Other possible worlds are, in their own terms, regular. These are worlds where things go exactly as they ought. Heat a dry piece of wood in a regular world and it will catch fire, if it is combustible there. If the wood is instead soluble in water, then when thrown into water it will dissolve. In a regular world where iron filings will cling to a magnet, iron filings do cling to magnets held close. In a regular world where magnets instead attract grass clippings, it is these which stick.

Irregularity is a relational characteristic of worlds. Just as one is a parent in virtue of his offspring, and a cup is overflowing in virtue of whatever is flowing out of it, a world is irregular in virtue of some condition or other. It is irregular in virtue of a given condition if some other condition does not prevail which ought to, given that the first condition does. For example, in the actual world it is true that, if a bar is a magnet, then provided conditions are suitable iron filings will cling to it. Therefore, the actual world is irregular, in virtue of the fact that a given bar is a magnet, if iron filings will not cling to it, or if some other thing does not happen which would have, had conditions been suitable for the clinging of iron filings.

The actual world is then irregular, at least, regarding the clinging of iron filings. Perhaps not regarding other ends. For example, in the actual world it is also true that, if a bar is a magnet, then provided conditions are suitable for the rusting of iron, it will rust. Suppose this magnet rusts and everything else is as it will be, provided conditions are suitable for the rusting of iron. Then regarding the rusting of iron, it is not so that the actual world is irregular in virtue of the fact that this bar is a magnet.

Say that in virtue of a true sentence A, a world is irregular regarding an end e if some sentence C is false there, while C ought to hold insofar as $A \triangleright_e C$ holds true. More concisely, say that *in virtue of A, w is irregular regarding e* if for some sentence C, $(A \triangleright_e C) \rightarrow (A \rightarrow C)$ is false at w. Say that *w is irregular in virtue of A* if, in virtue of A, w is irregular regarding some end. Finally, say that a world is *irregular* if it is irregular in virtue of some sentence; otherwise, say that it is *regular*.

In the distinction between regular and irregular possibilities there is a simple notion of comparative regularity of worlds. Now, fainthearted conditionals make claims about how things ought to go. Regular worlds are ones where things go as they ought. It might be thought that fainthearted conditionals make claims about comparative regularity, or that comparative regularity is one of the things that make fainthearted conditionals true.

For a proper understanding it is important to realize that this is not so. Indeed, there would otherwise be a circularity right at the center of my account of reasonable expectations. Whether a world is regular depends on which conditionals are true there, so to gain an understanding of comparative regularity you need an understanding of the truth conditions of fainthearted conditionals. If these truth conditions, in turn, presupposed comparative regularity, then to gain an understanding of them you would already need to understand that. This circularity would leave my account up in the air: neither the truth conditions nor the notion of comparative regularity would be sufficiently grounded in things independently understood. My account does not contain this circularity, however, since comparative regularity is not involved in spelling out the truth conditions of conditionals. Rather, it is involved in drawing the line between reasonable and unreasonable arguments.

Comparative similarity, on the other hand, *is* involved in these truth conditions. One formal difference between this notion and comparative regularity ties in with this. The former is a contingent relation. Which of two ways-things-might-have-been is most like the way things actually are depends, naturally, on the way things actually are. But comparative regularity is not a contingent relation. Which of two possibilities is more regular does not depend on contingent matters of fact. This difference between comparative similarity and regularity corresponds directly with a telling difference between conditional connectives and forms of argument in general. Conditionals are contingent, insofar as their truth and assertibility can vary from time to time and from place to place. But whether an argument is acceptable does not depend on what time it is, or where you happen to be.

Now I can state the critical policy more precisely. Fix a set of premises, and think of them as all of one's beliefs about some domain, so that the worlds where they hold true are just one's doxastic possibilities. As a first approximation, I shall say that an irregular world where these premises hold is *unnecessarily irregular*, regarding the premises, if there is a regular alternative to it—if, in other words, there is another premise world that *is* regular.

Ignoring all unnecessarily irregular counterexamples gives rise to what I shall call *allowed arguments*. An argument is valid if it has no counterexamples at all, but allowed if any counterexamples are unnecessarily irregular. I shall say that regular premise worlds *disqualify* unnecessarily irregular counterexamples, and that the premises of an allowed argument *allow* the conclusion.

Clearly, any valid argument is allowed. All counterexamples are unnecessarily irregular for the trivial reason that there are none. But, in addition, some invalid arguments are allowed. An invalid argument is allowed if some regular premise world disqualifies all of its counterexamples. Equivalently, an invalid argument is allowed if its conclusion is true at all regular premise worlds, and there is at least one of these.[11]

Allowed arguments add, to the logical consequences of what we know, anything which follows on assuming the world is regular. But they do so only if this assumption is compatible with what we know; where it is incompatible, they add nothing. Now, any child knows enough to exclude the possibility that the world is regular, so realistic arguments, which include in their premises a good bit of what we know about things, will rarely be allowed if they are not also valid. Odd that regular worlds and allowed arguments should throw any light, say, on prediction and explanation in science.

Odder still that regular worlds should appear in the role in which I have cast them, as more likely than their irregular alternatives, or as possibilities we normally reckon with, since regularity is quite as strange and unrealistic a notion as you might already think. For one thing, regularity is incompatible with *opposed* dispositions. In a regular world, nothing tends to behave in a particular way under a given condition, and at the same time tends to behave in an incompatible way under another condition, while both conditions prevail. The opposition of dispositions is common, however. I suppose it is even physically necessary.

Never mind any of this. I can make the notion of allowed argument more sophisticated in several ways. For a start, I can allow possibilities to be regular within some more-or-less restricted set of sentences paired with ends—call it the reasoner's *focus of attention*—which are of special interest. Irregularity will count only in virtue of sentences and ends that are in focus, so possibilities stand a better chance of being regular as far as any given focus of attention is concerned.[22]

---

[11] A reader pointed out a difficulty that arises if possibilities are compared in this way as to their regularity. Given a fair coin, you might say:

> If this coin is tossed 1000 times, then normally it will land heads approximately 500 times.

Now, 501 is approximately 500, so a world where the coin is tossed 1000 times and lands heads 501 of them is as regular as a world, otherwise similar, in which the coin is tossed and lands heads exactly 500 times. But then what about 502, 503,...1000 times?

Certainly, 501 times is approximately 500; certainly, 1000 times is not. As for the numbers in between, one just cannot say with the same certainty. Does the coin lands heads approximately 500 times when it lands so, say, 519 times? That depends on what you mean by 'approximately 500'.

The relation of allowed argument inherits the vagueness of such expressions. Suppose the coin, in the course of being tossed 1000 times, has already landed heads 501 times. Arguing as in later examples, the above conditional together with:

> The coin will land heads at least 501 times.

certainly allows the conclusion that the coin will land heads approximately 500 times. A little later though, the coin has already landed heads 519 times. With the additional premise:

> The coin will land heads at least 519 times.

is this conclusion still allowed? One just cannot say with the same certainty. That depends on what you mean by 'approximately 500'.

In addition, in section III, I shall develop the idea that some irregular possibilities are more regular than others. Comparative regularity will become a more discriminating notion, enabling irregular possibilities to be compared. This will make the critical policy less heavy-handed, since to disqualify a counterexample a fully regular premise world will no longer be required. Instead a more regular world will do—a world, roughly speaking, which is more as it ought to be.

Things are more transparent with the simpler notions though, and it is worth seeing how even the simple-minded critical policy accounts for reasonable expectations. Take the explanatory inference with premises:

> This bar is a magnet.
> If this bar is a magnet, then provided conditions are suitable iron filings will cling to it.

The second premise concerns conditions suitable, say, for *c*: the clinging of iron filings to a bar. I shall write them $M, M \rhd_c C$. Some worlds where these premises are true are irregular in virtue of M, 'this bar is a magnet'. These are worlds, say, where iron filings drop straight off the magnet because of a tremendous gravitational field. On the other hand, the premises surely do not *entail* irregularity; there must be a way to satisfy them while having things act as they tend to. Assuming for the moment that this is so, the critical policy allows the conclusion, C, which is beyond the reach of logic: 'iron filings will cling to this bar'. The argument with this conclusion is allowed, though invalid, since no world qualifies as a counterexample where the premises are true, but the conclusion is false. Any such world is irregular regarding the first premise, and is therefore disqualified by the regular premise world we have assumed.

I still have to show that the premises do not entail irregularity. I could use a modal frame for this but the following fact, which makes the simplest examples straightforward, will also do.[23] Let S be a sentence, in which either or both of $\rhd$ and $>$ appear. Let $[{}^{>}/_{\rhd}]S$ be the result of replacing each occurrence of $\rhd$ in S with an occurrence of $>$. Then S holds true in some regular world provided $[{}^{>}/_{\rhd}]S$ holds true in any world at all. It follows directly that the above premises do not entail irregularity. They hold true in some regular world since $M, M > C$ hold true in some world.[34]

---

[22] Whenever I restrict the focus of attention, I shall focus on the antecedents of those fainthearted conditionals appearing as premises, paired with the ends or regards with which they appear.

[23] To show that something is possible, I demonstrate that it is logically possible. My arguments assume that each satisfiable sentence expresses a real possibility.

For another example, consider premises:

> Either this bar is a magnet or it is a piece of wood covered with glue.
> If this bar is a magnet, then provided conditions are suitable iron filings will cling to it.
> If this bar is wood covered with glue, then provided conditions are suitable iron filings will cling to it.

Letting W be 'this bar is a piece of wood covered with glue', they can be written $M \vee W$, $M \rhd_{\epsilon} C$, $W \rhd_{\epsilon} C$. By the earlier observation, they hold true in some regular world. It follows that the conclusion C is allowed, since any premise world is unnecessarily irregular in which C is false. It is irregular in virtue of M, if M holds true there, or in virtue of W, if instead W holds true. This irregularity is unnecessary since there is a regular alternative.

Characteristically, the relation is not a simple one between evidence and the expectations it renders reasonable. What once seemed likely can turn out to be untrue, and then it no longer seems likely at all. Suppose we have not actually *seen* iron filings clinging to a given bar, which we take to be a magnet, but we expect that they will. Then we check and find that they do not. This observation is compatible with magnetic theory and the supposition that the bar is indeed a magnet; it is the sort of thing that you can expect if there is interference from some other force. For this reason, the observation simply adds to the suppositions that led to the expectation, rather than calling for their revision. With this addition, however, there is no longer a basis for the expectation that iron filings will cling.

We have seen that allowed arguments can take us, beyond the logical consequences of premises, to reasonable expectations. Now, we shall see that allowed arguments are *defeasible*; unlike logical consequences, reasonable expectations need not accumulate as the evidence on which they are based does. To see this, notice that the premises of both of the previous arguments entail irregularity, once the observation $\neg C$ is added. Therefore, since there is no longer a regular premise world, the counterexamples to the conclusion C, though irregular, are no longer unnecessarily so. It follows that the conclusion C is no longer allowed.

---

[34] The observation can be verified as follows. An interpretation with a world w making $[{}^{>}/_{\rhd}]S$ true can be turned into an interpretation in which w is regular, and in which w makes S true. To do this simply stipulate, for each world x, proposition p and end ϵ, that $\wp(x,p,\epsilon)$ is the vacuous proposition, containing all possible worlds. (Then the conditions implicit in provisos are all vacuous.) This collapses the truth conditions of $\rhd$ into those of $>$, so that any sentence T holds true, at any given world x, just in case $[{}^{>}/_{\rhd}]T$ does. In particular, S holds true at w, since by assumption $[{}^{>}/_{\rhd}]S$ holds true there. In addition, w is regular. Every sentence of the form $(A \rhd_{\epsilon} C) \rightarrow (A \rightarrow C)$ must hold there, since modus ponens is valid in the logic of $>$.

The previous example shows that finding its conclusion to be false defeats an allowed argument. The next example shows that merely finding reason to *expect* the conclusion is false can have the same effect. Consider premises:

> This engine is warm.
> If this engine is warm, then under normal circumstances it will start easily.

The second premise concerns circumstances normal with regard to *e*: the ease with which the engine will start. I shall write them W, $W \rhd_e E$. As in the first example, they allow but do not entail E. Consider in addition:

> This engine is flooded.
> If this engine is flooded, then under normal circumstances it will not start easily.

This conditional also concerns circumstances normal regarding the ease with which the engine will start. I shall write these additional premises F, $F \rhd_e \neg E$, and on their own they allow but do not entail $\neg E$. With all four premises, however, neither conclusion is allowed. First, two counterexamples show that the combined premises entail neither conclusion. World u of the following interpretation is a counterexample to the conclusion E:

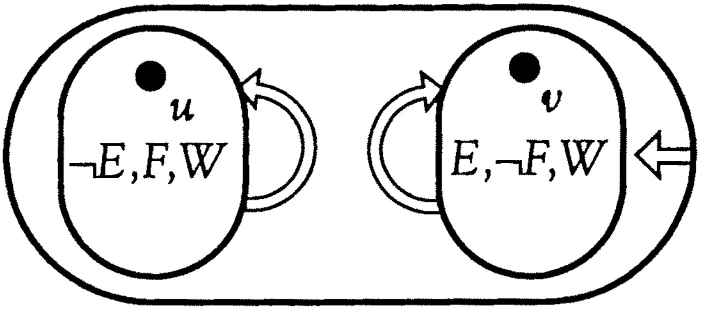


Figure 2

This frame provides a counterexample to the conclusion $\neg E$, too. These counterexamples cannot be put aside as unnecessarily irregular since *every* premise world is irregular. A premise world is irregular in virtue of F if E is true there, but irregular in virtue of w if E is false.

III. DEGREES OF REGULARITY

This picture of the relation between fainthearted conditionals and reasonable expectations is a sketch and there remains a lot to fill in. To be realistic, what we know is not compatible with the assumption that the world is regular. The policy of assuming regularity, where possible, needs to become a policy of assuming *as much* regularity as possible. In fact, it is straightforward to elaborate it in this way.

It can happen that a world is irregular in virtue of one sentence but not in virtue of another. A world can be irregular regarding one end, but not regarding another. Thus, worlds can be compared in their degree of regularity: say that one world is *as regular as* another if for any sentence in virtue of which the first world is irregular regarding some end, the second world, too, is irregular in virtue of that sentence, regarding that end. Say that one world is *more regular than* another if it is as regular as the other, but not the other way around. There is a sentence in virtue of which the less regular world is irregular regarding some end, while the more regular world is not irregular in virtue of this sentence, regarding this end.

These notions can once again be made relative to a set of sentences paired with ends, the focus of attention. Thus, one world is *as regular as* another, *as far as the focus is concerned*, if for any focused sentence and end in virtue of which the first world is irregular, the second, too, is irregular in virtue of that sentence, regarding that end.

With this refinement I can make the critical policy more discriminating. Instead of requiring a fully regular premise world to disqualify a counterexample, from now on a more regular one will do. Say that a counterexample to an argument is *unnecessarily* irregular if there is some more regular premise world. As before, say that an argument is *allowed* if any counterexamples are unnecessarily irregular.

The more discriminating notion of allowed argument leaves the analyses of the examples in the previous section as they were. There, to show that an argument is allowed, I showed that its premises do not entail irregularity. Each demonstration relied on a regular premise world that disqualified, as counterexamples, all irregular premise worlds. Now, since these regular worlds are more regular than any irregular world, they still disqualify the counterexamples and the analyses remain intact. But we can, in addition, look at some more interesting arguments, whose premises do entail irregularity.

You have before you a dry piece of wood and a lump of sugar. Wood is combustible and sugar is soluble in water. You heat the wood through and through, but it will not burn; there is not enough oxygen, say. Still you can expect the sugar to dissolve, on placing it in water. We have premises:

> The wood is thoroughly heated.
> The sugar is placed in water.
> The wood will not burn.
> If it is heated thoroughly, then provided conditions are suitable the wood will burn.
> If it is placed in water, then provided conditions are suitable the sugar will dissolve.

The first conditional concerns conditions suitable for *c*, the combustion of wood; the second, conditions suitable for *d*, the dissolving of sugar. Write them H, S, $\neg B$, $H \rhd_c B$, $S \rhd_d D$. To keep things simple, focus attention on H and S, the antecedents of the two conditional premises, paired respectively with *c* and *d*.

With the simpleminded policy on counterexamples, these premises do not allow the conclusion, D, 'the sugar will dissolve'. The premises entail irregularity in virtue of H, so premise worlds where the sugar does not dissolve will qualify as counterexamples. (An interpretation based on the example frame shows that there are some.) On the other hand, the following interpretation shows the premises are compatible with regularity as far as S is concerned. There is a premise world where the sugar is placed in water and everything goes as it ought to under conditions suitable for dissolving sugar.

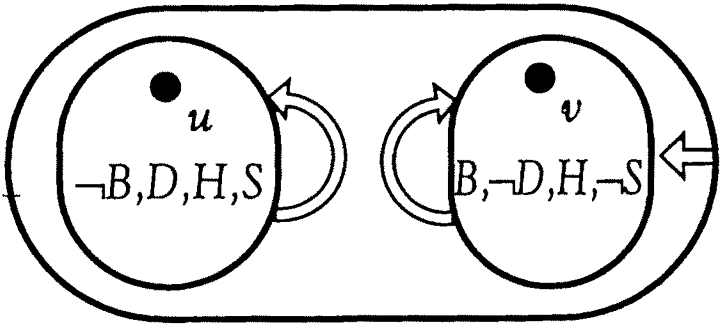


Figure 3

It is straightforward to show that the premises are true at u, as is every sentence of the form $(S \rhd_d T) \rightarrow (S \rightarrow T)$. Therefore, u is not irregular in virtue of S, regarding d. In consequence, u is more regular, relative to our focus of attention, than any premise world where D is false. For while u is irregular only in virtue of H (regarding c), any such world is irregular in virtue both of H (regarding c) and of S (regarding d). Therefore, no world where D is false qualifies as a counterexample to any argument leading from these premises. The conclusion, 'the sugar will dissolve', is allowed.

Magnets attract iron filings; also, magnets rust. Suppose that this bar is a magnet but it will not rust. There is not enough oxygen in the surrounding atmosphere, say. Then it is still reasonable to expect that iron filings will cling. We have premises:

> This bar is a magnet.
> This bar will not rust.
> If this bar is a magnet, then provided conditions are suitable iron filings will cling to it.
> If this bar is a magnet, then provided conditions are suitable it will rust.

Letting r be the rusting of iron, I shall write them $M \rhd_c C$, $M \rhd_r R$, M, $\neg R$. To keep things simple, restrict the focus of attention to M paired with c, and M paired with r. Clearly, the premises entail irregularity in virtue of M, regarding r. But they do not entail irregularity in virtue of M regarding c. There is a premise world where for any sentence S, $(M \rhd_c S) \rightarrow (M \rightarrow S)$ holds true.[25] It follows that the premises allow the conclusion C, 'iron filings will cling to this bar'. Any counterexample to this argument is irregular in virtue of M, not only regarding r, but also regarding c. Clearly, any such counterexample is unnecessarily irregular.

Here is a more interesting case.[26] Consider premises:

---

[25] This can be shown by modifying the example frame slightly. As far as r is concerned, let $\wp$ be just as in the example frame. But as far as c is concerned, let $\wp$ be the identity function: for each world x, and for each proposition p, $\wp(x,p,c) = p$. Further, let C and M be true at both u and v, but let R be true only at u. Now u is a suitable world.

[26] It was suggested in schematic form by a reviewer. Renata Tinini then came up with this particular example, which Richard Feynman once put to the test. It seems he distractedly asked for both lemon and milk in his tea. The response became the title of his autobiography: *Surely You're Joking, Mr. Feynman* (New York: Norton, 1985).

> If there is milk in the tea, then normally it will taste good.
> If there is lemon juice in the tea, then normally it will taste good.
> If there is milk and lemon juice in the tea, then normally it will not taste good.
> There is milk and lemon juice in the tea.

I shall assume that the premises concern circumstances normal regarding one and the same thing: the taste of the tea, t. I shall write them $M \rhd_t G$, $L \rhd_t G$, $(M\&L) \rhd_t \neg G$, $M\&L$.

It seems reasonable to expect $\neg G$, "the tea will not taste good." But, restricting the focus of attention to M, L, and $M\&L$, the antecedents of the conditional premises, paired with t, the premises do not allow this conclusion.[17] As you might expect—and as a model construction shows—some premise worlds are irregular in virtue of $M\&L$, but not irregular in virtue of M or in virtue of L. At these worlds G holds true, so they are counterexamples to the conclusion $\neg G$. Furthermore, there is no more regular premise world to disqualify them. A more regular world would be irregular in virtue of *none* of M, L, and $M\&L$. But every premise world is irregular in virtue of one or more of these sentences, accordingly as G or $\neg G$ is true there.

It might seem that allowed argument cannot account for reasonable expectations in this case. But that would be too quick. Tea with milk normally does not contain lemon juice. Tea with lemon juice normally does not contain milk. Add the missing premises $M \rhd_t \neg L$ and $L \rhd_t \neg M$ and things are different. Then *all* premise worlds are irregular in virtue of both M and L. But, as I shall show in a minute, not all are irregular in virtue of $M\&L$. It follows that any premise world where G holds, being irregular in virtue of $M\&L$, is unnecessarily irregular. So the conclusion $\neg G$ is allowed.

I still have to show the premises hold true in some world that is not irregular in virtue of $M\&L$. Here is a suitable interpretation. There are three possible worlds: u, v, and w. Where x is any of these worlds, $\wp(x,\{v\},t) = \{v\}$, while for any p other than $\{v\}$, $\wp(x,p,t) = p \cap \{u,w\}$. Finally, M and G are true at u, but L is false; M and L are true at v, but G is false; and L and G are true at w, but M is false. As before, it does not matter which comparative-similarity relation the interpretation has.

---

[17] Since the conditional premises all concern circumstances that are normal in the same regard, I can suppress t in the rest of the discussion.

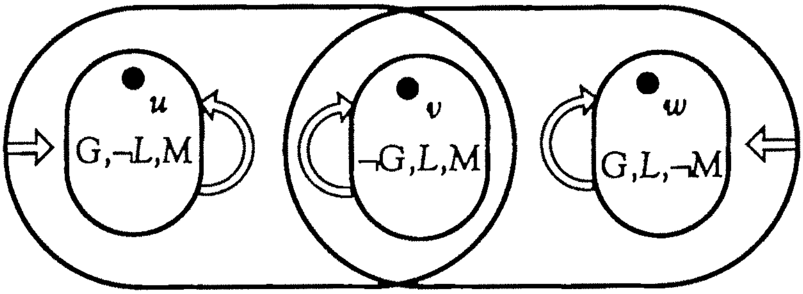


Figure 4

It is now a question of routine to check that the premises are true at v (including the two "hidden" ones), and that v makes true $((M\&L) \rhd C) \to ((M\&L) \to C)$ for any sentence C, and is therefore not irregular in virtue of $M\&L$.

MICHAEL MORREAU

University of Maryland/College Park

BOOK REVIEWS

Liberation from Self: A Theory of Personal Autonomy. BERNARD BEROFSKY. New York: Cambridge University Press, 1996. ix+270p. Cloth $54.95.

In the past decade, the subject of autonomy has received the kind of attention and detailed investigation that other concepts such as free will and moral responsibility have long enjoyed. Various authors have presented their accounts of the nature of autonomy and of its value. Since the varying accounts are inconsistent with one another, many must be (partly) wrong—probably all of them! Recognizing this, some authors hedged their bets by denying that "their" conception of autonomy was the only one possible. Following the modesty of John Rawls who titled his book *A Theory of Justice*, they hoped to produce an account of a theory of autonomy; one that helped to think about some range of problems and illuminated some normative issues in which the concept of autonomy was implicated.

Bernard Berofsky's book is more ambitious. He claims to have the right one:

> ...[my] conception of autonomy as liberation is not just one conception among several viable ones, each meeting different desiderata, but is rather the one idea of personal autonomy that survives conceptual and empirical scrutiny (9).

According to Berofsky, there are three main components of the core notion of autonomy: independence, (instrumental) rationality, and personal integrity (9-10). Further understanding of these components leads to the additional elements of freedom and objectivity. Autonomy and freedom are distinct notions but some kind of positive freedom is necessary for autonomy (13). Spontaneity and competence, not reflective evaluation, are the "cornerstones" of autonomy. An autonomous person is "objective" in the sense that her responses track the world appropriately (14). There is no "self" who plays the role of author of the decisions and actions of an autonomous agent (14). Autonomy is, therefore, not best understood as self-determination but as liberation, that is, the transcendence and liberation of ourselves from the origins of ideas, emotions, and desires that do not track the world appropriately or correctly. But autonomy does not require subscription to various ideals, such as human flourishing or morality (11).

0022-362X/97/9404/212–16

Autonomy is a matter of degree and is grounded in the agent's relation to herself and the world. The account is nonhistorical, as it does not depend on the agent's endorsing the means by which her current will is determined. For those who demand their philosophy fit a T-shirt—autonomy is not the freedom to express our origins; it is the freedom from those origins (225).

In the course of making and defending these claims, Berofsky has many interesting and novel things to say. He is particularly good at bringing to bear various psychological views that have usually been ignored.

Any claim that a conception of autonomy is the only one that stands up to critical scrutiny assumes some way of limiting the domain. Otherwise, one could say: "Here is another idea of autonomy that survives conceptual and empirical scrutiny. A person is autonomous if and only if she does things of which her mother would have disapproved." This notion is nonempty; there are such people. It is not trivially satisfied (some people only do things their mother would have approved). It is a consistent, coherent notion that is testable in practice.

One way of delimiting the domain is by making a distinction between an abstract "concept" of autonomy and a substantive conception that is filled in differently by different philosophers. But Berofsky does not believe in this (9). Nor does he appeal to paradigm cases. While he starts out with the provisional assumption that autonomy is some kind of self-direction, he winds up explicitly claiming that "the assumption that autonomy is either self-origination or self-direction...must be abandoned" (1).

Another way of delimiting the domain is by listing a set of problems or issues that thinking about the concept in question is supposed to resolve. Are we interested in some deep, metaphysical issues about the nature of rationality and freedom as, for example, Immanuel Kant is? Are we interested in promoting some ideal of human development and fulfillment as J. S. Mill is? Does the interest lie in some very specific issue such as, say, freedom of expression? Or perhaps we want to use the concept to ground some broader view of, say, the legitimacy of the state? Are we interested in explaining what makes moral agents worthy of respect and distinguishes them from, say, animals?

It is only against such a background that we can evaluate a claim such as "autonomy is not the same as freedom" or that "Jehovah's Witnesses lack autonomy because they lack rationality" (117). Take the former claim about freedom. I would translate it as something like: as I define the concept of autonomy, for the purpose of thinking about moral responsibility, given the constraints on an analysis of autonomy which I set out earlier, it is useful to distinguish autonomy from another set of phenomena that have been traditionally called "being free." Failure to do so results in the following problems: for example, we cannot explain why Odysseus expressed his autonomy in asking that his fellow sailors restrain his freedom.

So, in my view, the proper form of criticism of any proposed theory is that the set of problems is uninteresting or confused or should be replaced by some other set of questions. Or that, although the problems are genuine, this way of looking at autonomy is not helpful or useful or insightful in thinking about these issues. And that could be for a number of different reasons varying from internal problems with the concept—it is self-contradictory or leads to some kind of regress difficulty—to the criticism that while the concept is clear enough and consistent, it runs too many things together or makes distinctions where they are not needed.

In light of this methodological thesis, let us look at a substantive criticism Berofsky makes of my views.[1] On any theory of autonomy, the following question can arise: Can a person who has autonomously decided to abandon her independent decision making still be counted as autonomous? On my view, the answer is positive. An individual who decides that her values and goals are best furthered by ceasing to reflect upon or question the commands of her guru, and whose decision is not made under conditions that are inimical to rational decision making, remains an autonomous person. Berofsky, by contrast:

> prefer[s] to say that, through the abandonment of procedural independence (I will not make decisions for myself) *and* substantive independence (my guru will), the agent has made an autonomous decision to become heteronomous (127).

The first thing to note is that Berofsky and I are using terms such as 'procedural' and 'substantive independence' in different ways. What I mean by the former is (roughly) the absence of causal influences that interfere with our thinking correctly: for example, hypnosis, brain washing, and the like. By the latter, I mean exercising a capacity for critical reflection. Berofsky means by the former (roughly) what I mean by substantive independence, that is, only accepting principles, desires, and the like, in a critical, reasoned fashion. By the latter, he means not turning over control of my actions to another person.

---

[1] *The Theory and Practice of Autonomy* (New York: Cambridge, 1988).

So when he concludes that "a truly substantively dependent agent has in fact *also* abandoned procedural independence" (127), this is analytic, given his definition of the terms. But all this is simply a matter of translating between the two vocabularies. We both agree that the person who becomes a Jehovah's Witness (Berofsky's example) is no longer thinking for herself (exercising her capacity for critical reflection). We both also agree that such a person still has the capacity; she has not had a lobotomy. We also, presumably, both agree that the decision reflects a fundamental commitment by the person as to how she wants to live her life, what kind of person she wants to be. But I claim what she has autonomously abandoned still leaves her an autonomous person, while Berofsky believes that she has autonomously abandoned her autonomy. Who is correct here?

I do not see any way of answering that question other than by asking which view provides better answers to various other questions. For example, just to consider some normative (as opposed to conceptual) questions: Is it a good thing to be autonomous? Why? How does autonomy ground (if it does) the claims of others to our moral attention?

As Berofsky notes, I believe that morality requires us to give weight to the way others define and value the world in deciding how we should act. Should I then respect the Jehovah's Witness? After all, says Berofsky, her wishes are not really hers (129). Berofsky is quite right that I would reject the claim that such a person has no claim to be respected as a person. But I would distinguish between two senses of her wishes not being her own. If one means that she has not independently reflected upon the command that she solicit door to door, then this is true. But, on the other hand, she has decided that she wants to define her life as a follower of the way. For her, the Lord's wishes *are* her commands. Therefore, in trying to act responsibly toward her, I must see that she attributes value to obedience. I must not interfere with that relationship (assuming it does not violate other moral demands). I certainly may think that her life is not a very good one, but it does have a claim on me.

What are the implications of Berofsky's view of the Jehovah's Witness as heteronomous? He does not want to be committed to the view that her life has no value, so he must reject the thesis that autonomy is a necessary condition of a life's having value. Indeed, he believes that the decision to become heteronomous can be a fully rational one. An autonomous person may be "barred from feeling the depth of community sentiment and the sense of belonging unique to [certain] communities" (213). Autonomy does not imply a commitment to morality or vice versa. But by now we are left wondering whether autonomy has any value.

Berofsky is quite right to reject strong views that claim autonomy as the supreme value or as a necessary condition of a life's having value, but it is difficult to see on his account why people have attributed such importance to autonomy. Why should we care whether the Amish have the right to keep their children out of the public school system or not? I suspect that those who identify autonomy with the capacity for critical examination, unlike those who locate autonomy at the level of exercise, think that there is more the state can, and ought to, do toward promoting autonomy via the system of public education.

Berofsky's book makes a definite contribution to the literature on autonomy. Had it more explicitly set out the kinds of problems for which we need a theory of autonomy, its contribution would have been enhanced. As Christine Korsgaard[2] says, with respect to opposing philosophical views: "The correct view is not going to be the one left standing when the contradictions and absurdities of all the others have finally been exposed. It is going to be the one that answers best to the human concerns which motivate the study of philosophy in the first place" (*ibid.*, p. xiii).

GERALD DWORKIN

University of Illinois/Chicago
University of California/Davis

---

[2] *Creating the Kingdom of Ends* (New York: Cambridge, 1996).

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 5, MAY 1997

PROBABILITY AS A GUIDE TO LIFE[*]

Let us assume that people's degrees of belief conform to the probability calculus. This is an idealization, but it will not matter for our purposes here. In line with normal usage, we shall refer to these degrees of belief as *subjective probabilities*. Then we can say that agents are *subjectively rational* if they choose those actions which their subjective probabilities imply are most likely to achieve what they want. More precisely, given that agents normally attach different utilities to a number of different ends (another useful idealization), subjectively rational agents will choose those actions which will maximize their subjective expected utility.

An example will be useful for what follows. Suppose Alma offers you her £2 to your £1, with you to win all the money if a coin lands heads. Should you accept or decline the bet? If you have no aversion to betting, and your utility function varies linearly with money for these small amounts, then subjective rationality dictates that you should accept her bet if your subjective probability for heads is greater than 1/3 (for example, if you think the coin fair), and you should decline if it is less than 1/3.

Note that in this example the possible outcomes (heads/tails) are causally and probabilistically independent of the choice of actions (accept/decline). We shall concentrate on examples like this throughout, in order to bypass the debate between causal and evidential decision theory. A fully general account of subjective rationality needs to specify whether decisions should depend on the degrees to which an agent believes that certain actions will *cause* cer-

---

[*] We would like to thank Dorothy Edgington, Peter Milne, and especially Scott Sturgeon for comments on earlier versions of this paper.

0022-362X/97/9405/217–43

tain outcomes, rather than on the agent's conditional degrees of belief in those outcomes given those actions. This is an important issue, but it cuts across our present concerns: the points we make will apply equally to the degrees of belief invoked by causal and by evidential decision theories. (In the examples on which we shall focus, where outcomes are independent of actions, both theories agree that choices should depend on agents' simple degrees of belief in outcomes.)

I. OBJECTIVE CORRECTNESS IN DEGREES OF BELIEF

Whatever view is taken on the causal-evidential issue, there is clearly little point in being subjectively rational if your degrees of belief are objectively inaccurate.[1] Consider Betty, who has normal nonprobabilistic beliefs about standard coins, yet believes to degree 0.9 that such coins will land heads when tossed. Betty will put up large sums of money against small ones on any ordinary coin's coming down heads. We take it that this is a bad thing for Betty to do. Yet Betty is still subjectively rational, since she chooses those actions which maximize her subjective expected utility.

Betty is subjectively rational. Moreover, she need not have any false beliefs, since believing to degree 0.9 that the coin will land heads is not the same as (fully) believing that in some objective sense its probability of heads is 0.9, and we can suppose Betty does not believe this. Nevertheless, there is something objectively unsatisfactory about Betty's degrees of belief. They will lead her to act in a way that is subjectively rational yet objectively inadvisable. We shall investigate objective desiderata on subjective probabilities, and consider what can be said about which degrees of belief are objectively *correct* as a basis for decision. After exploring various possible principles governing objective desiderata on degrees of belief, we shall eventually conclude that the correct degree of belief in an outcome is not necessarily equal to its single-case probability, but rather to its *relative probability*, by which we mean the *objective probability* of that outcome *relative* to the features of the situation which the agent knows about.

Let us stress immediately that we are concerned here with the correctness of degrees of belief from a *prudential* point of view. Our question is: Which are the right degrees of belief to act on, given that you want such-and-such results? Other philosophers start with a more epistemological question: Which degrees of belief ought you to adopt, given that you have such-and-such information that bears on the occurrence of the outcome?[2] We are interested here in the first, prudential question. Moreover, we take this question to be different from the evidential question. Degrees of belief are made prudentially correct by objective features of the world, and are correct in this sense even for agents who are in no epistemological position to adopt those degrees of belief. Of course, only those agents who are fully informed about the relevant objective features of the world can *ensure* they have the correct degrees of belief. But, even so, you do not need to have such knowledge in order for the relevant notion of correctness to apply to you. It would be correct, in our sense, for you to give up smoking, if you want to avoid cancer, even if you have never heard about the smoking-cancer link.

---

[1] Cf. D.H. Mellor, "Chance and Degrees of Belief," in R. McLaughlin, ed., *What? Where? When? Why?* (Dordrecht: Reidel, 1982), pp. 49-68.

This last point marks an important difference between our approach and that of the tradition of writers on "direct inference," such as Hans Reichenbach, H. E. Kyburg, Isaac Levi,[3] and J. L. Pollock (*op. cit.*). These writers, like us, address the question of which degrees of belief agents ought to adopt toward outcomes. They are in agreement with our eventual conclusion, at least to the extent that they hold that agents ought to set their degrees of belief equal to the relevant relative probabilities when they are in an epistemological position to do so. It is not clear, however, that writers in this tradition share our central concern with prudential rationality. Writers on direct inference tend simply to take it for granted that you ought to set your degrees of belief equal to the relative probabilities if you can. Their real focus is rather on the evidential issues that arise when you cannot do this, because you do not know the requisite objective probabilities, but only have incomplete evidence about them. By contrast, according to our sense of "correctness," the degrees of belief of an agent who has no knowledge of the relevant objective probabilities are still constrained by those objective probabilities, in the sense that the correct degrees of belief for the agent to act on are still those which match the objective probabilities. Evidential issues that arise for agents who lack full knowledge of the relevant objective probabilities are important, but they lie beyond the scope of this paper. Our concern is to show that certain objective features of the world—namely, relative probabilities—make it prudentially correct to adopt corresponding degrees of belief. How far agents can achieve this, and what they should do if their lack of information hampers them, are further questions on which we shall comment only in passing.

---

[2] For this distinction, cf. J.L. Pollock, *Nomic Probability and the Foundations of Induction* (New York: Oxford, 1990), pp. 22-23.

[3] See, for example, Reichenbach, *The Theory of Probability* (Berkeley: California UP, 1949); Kyburg, *The Logical Foundations of Statistical Inference* (Dordrecht: Reidel, 1974); Levi, "Direct Inference," in this JOURNAL, LXXIV, 1 (January 1977): 5-29.

II. THE PERFECT PRINCIPLE

Our aim, then, is to specify what makes certain degrees of belief correct from a prudential point of view. The first idea we want briefly to consider is the principle that the correct degrees of belief are those which match the truth. According to this principle, if outcome o_k is going to occur, then the right degree of belief in o_k is 1, and if o_k is not going to occur, the right degree of belief in o_k is 0. Thus, we would have:

(1) The *perfect principle*: If the k^{th} of mutually exclusive outcomes $\{o_i\}$ is going to occur, then the correct subjective probability for o_i is 1 for $i=k$, and 0 otherwise.

We take it, however, that this is not a satisfactory account of which degrees of belief are objectively correct from a prudential point of view. If determinism is false, it is not in general fixed at the time of decision what the outcome will be. So even an agent who is omniscient about the laws of nature and current state of the universe would not be able to implement the perfect principle. In many decision situations, then, the prudentially correct degrees of belief must be fixed by the objective probabilities of outcomes rather than by what is actually going to happen.

III. THE OBJECTIVE PRINCIPLE

The obvious alternative to the perfect principle is to say that the correct degrees of belief are those which match the objective probabilities:

(2) The *objective principle*: the correct subjective probabilities for outcomes $\{o_i\}$ are those which match the objective probabilities of those outcomes.

Note that agents who have the right degrees of belief in the sense of principle (2) and are also subjectively rational[4] will act in such a way as to maximize not just subjective expected utility, but also objective expected utility.

---

[4] We shall take this qualification about subjective rationality as read henceforth. Note that this warrants our using phrases like 'setting your degrees of belief equal to such-and-such objective probabilities' interchangeably with 'acting on such-and-such objective probabilities'.

The obvious question about the objective principle is how to understand its reference to "objective probabilities." This will be our focus here. We shall argue that the best interpretation of the objective principle requires us to read 'objective probability', not in the most obvious sense of *single-case probability*, but rather as *probability relative to a description* of the decision situation.

IV. SINGLE-CASE PROBABILITIES

The nature of single-case probabilities will be discussed further in what follows. For the moment, let us simply think of them as the rock-bottom metaphysical probabilities fixed by all probabilistic laws and all particular facts about specific situations. Quantum events provide the paradigm. The 0.5 probability of a radium atom decaying by A.D. 3615 is a single-case probability, in that there are no further contemporary facts about the radium atom or its surroundings that imply that the probability of its decaying by A.D. 3615 is different from 0.5.

Note that, if determinism were true, then single-case probabilities would never be different from 0 or 1. Nondegenerate single-case probabilities require outcomes whose occurrence or nonoccurrence is not fixed by present circumstances and laws.

Note also that such nondegenerate single-case probabilities change over time. This happens when circumstances relevant to the outcome, but not themselves previously determined, either occur or fail to occur. Thus, the single-case probability of a given radium atom decaying by a particular time will decrease as time passes and the atom does not decay, but will suddenly jump to one if the atom does decay. To take a more everyday example, due to David Lewis,[4] imagine people who move through a multipath, multi-exit maze, deciding at each branch point whether to turn left or right by some quantum-mechanically random procedure. When they start, their reaching a given exit on the other side of the maze will have a given single-case probability different from 0 or 1. But this single-case probability will increase or decrease, depending on which choices they actually make as they pass through the maze, until it is determined where they will come out, by which time the single-case probability will have become 1 or 0.

V. THE SINGLE-CASE PRINCIPLE

One obvious way to read 'objective probabilities' in principle (2) is as the single-case probabilities at the time of the agent's decision. If you are betting on where someone will come out of the maze, do you not ideally want to know the *current* single-case probability of their emerging from each of the various exits? This suggests the following principle:

---

[4] "A Subjectivist's Guide to Objective Chance," in R.C. Jeffrey, ed., *Studies in Inductive Logic and Probability, Volume II* (Berkeley: California UP, 1980).

(3) The *single-case principle*: the correct subjective probabilities for an agent to attach to outcomes $\{o_i\}$ are equal to the single-case probabilities of the $\{o_i\}$ at the time of the agent's decision.

Note that this single-case principle is equivalent to the perfect principle in all those cases where there is any practical possibility of humans using the perfect principle. For we can only now know whether some future outcome is going to occur if that outcome is already determined; and in such cases, the single-case probability of that outcome will already be 0 or 1.

It may seem obvious that, if there is a right and a wrong about which degrees of belief you ought to act on, then it must be as specified by the single-case principle. For the only way to attach a unique number to the objective probability of an outcome at the time of some decision is to equate it with the single-case probability at that time. So it seems to follow that acting on the single-case probability is the only way to guarantee that you will always choose the action that will maximize your objective expected utility.[6]

Nevertheless, despite this plausible line of reasoning, we think that the single-case principle is not the right way to link degrees of belief with objective probabilities. We think there are many situations where the objectively correct degrees of belief do not match single-case probabilities. We shall formulate a different principle, which we shall call the *relative principle*, which will cover these cases as well. This more general principle will assert that, in general, the correct degrees of belief do not match single-case probabilities, but rather a kind of probability we shall call *relative probability* (because it is relative to a description of the set-up facing the agent).

The relation between our preferred principle, the relative principle, and the single-case principle is analogous to that between the single-case principle and the perfect principle. We earlier rejected the perfect principle in favor of the single-case principle on the grounds that the perfect principle does not cover all cases of prudentially correct choices. We reject the single-case principle for the same reason. There are prudentially correct choices beyond those recommended by the single-case principle.

---

[6] Of course, degrees of belief other than those which match single-case probabilities can lead to the right choice in *particular* decision situations. The point of the single-case principle is rather that it ensures maximization of single-case expected utility in *every* decision situation. (Analogously, in nonprobabilistic contexts, while a choice of means recommended as effective by false beliefs will *sometimes* lead to desired results, those so recommended by true beliefs will *always* do so. Cf. Papineau, *Philosophical Naturalism* (Cambridge: Blackwell, 1993), ch. 3.)

In what follows, we shall first illustrate the relative principle with examples and explain how it works. We hope these examples will help make the relative principle plausible. But our ultimate rationale for favoring this principle over the single-case principle will be that it turns out to yield a simpler overall theory of prudential rationality. By the time we reach the end here, we shall be able to show that the relative principle yields a uniform and integrated account of prudential rationality, while the single-case principle does not.

Although we oppose the single-case principle, we suspect that it captures many people's intuitive understanding of prudentially correct degrees of belief. Even if this suspicion is right, however, it is difficult to find any straightforward expression of the single-case principle in the philosophical literature. Those authors who come closest to arguing that the right degrees of belief must match the single-case probabilities do not generally adopt the simple version of the single-case principle given above. Instead, they defend more complicated formulations, designed in part to block some of the difficulties we are about to raise. Thus, D. H. Mellor, Wesley Salmon,[7] and Lewis (*op. cit.*) all forge a central connection between degrees of belief and single-case probabilities, but in each case via something more complicated than the simple single-case principle given above. We shall comment on these writers in due course. But for the moment, we shall retain the single-case principle as our target. This is because our disagreement is not on matters of detail, but rather on the whole idea that the objectively correct degrees of belief are those which match the single-case probabilities. It will be easier to make our objections clear if we start with a simple version of this idea.

VI. RELATIVE PROBABILITIES

It is high time we explained what we mean by the relative principle.

(4) The *relative principle*: the correct subjective probabilities for an agent to attach to outcomes $\{o_i\}$ are the probabilities of the $\{o_i\}$ relative to the agent's knowledge of the set-up.

To illustrate this principle, and the associated notion of relative probability, consider this example. Alma tosses a fair coin and places her hand over it as it lands. We shall say that the relative probability of finding heads when Alma's hand is removed, *in relation to the description just given of this situation*, is 0.5. Note that this probability of 0.5 is not a single-case probability, since the single-case probability of heads is already either 0 or 1. Yet it is clearly the right probability to act on if you have not seen under Alma's hand, in line with the relative principle. For example, it is the right probability for working out that you should accept the bet we earlier described Alma as offering, since this 0.5 probability implies that the bet offers an expected gain of 50p.

---

[7] Mellor, "Chance and Degrees of Belief," and *The Matter of Chance* (New York: Cambridge, 1971); and Salmon, "Dynamic Rationality: Propensity, Probability and Credence," in J. Fetzer, ed., *Probability and Causality* (Dordrecht: Reidel, 1988).

It is not essential to the point we are making that the relevant result already has occurred. Supposed Alma tosses the coin and then offers her bet when it is in the air. Let us assume, as seems plausible, that the result, though still in the future, is by now determined. So the single-case probability of heads is already 1 or 0. Yet the relative probability of 0.5 is still clearly the probability on which you ought to act.

Nor is it essential to the point that the single-case probability of the relevant outcome already be 1 or 0. Suppose that Alma has a machine that makes one of two kinds of devices at random. On each run of the machine, a radioactive source decides with probability 0.5 whether the machine will make device *A* or device *B*. Each device will then print out 'heads' or 'tails', again on a quantum-mechanical basis. Device *A* gives 'heads' a single-case probability of 0.8. Device *B* gives 'heads' a single-case probability of 0.2. You know that the device in front of you is made by the machine, but you do not know whether it is an *A* or a *B*. Again, it seems obvious that you ought to decide whether to accept Alma's bet on the basis of the relative probability of 0.5 for 'heads', even though the single-case probability is already 0.8 or 0.2.

VII. PROBABILISTIC LAWS

It will be helpful to be more explicit about relative probabilities. For this we need the notion of a *probabilistic* law. We shall take the following as the basic form of such a law: 'The probability of an *A* being a *B* is *p*' ('The probability of a tossed coin concealed under a hand turning out to be heads is 0.5'). We take it that there are many objective truths of this form. Moreover, we take it that many of these are known to be true, on the basis of inferences from statistical data about the proportions of *B*s observed in classes of *A*s.

Given a particular situation that satisfies description *D* and a particular exemplification *o* of type *O*, the probability of *o* relative to *D* is *p* if and only if there is a probabilistic law 'The probability of a *D* being an *O* is *p*'. Note that these relative probabilities are like single-case probabilities in that they attach to a particular outcome (a particular exemplification of a type) rather than to the type itself. But they do not attach to it simpliciter, but rather to the pair of the outcome and some description of the decision situation.

Single-case probabilities can be thought of as a special case of relative probabilities, namely, as the relative probabilities fixed by *all* the relevant features of the situation. If we think of single-case probabilities in this way, then the single-case probability at *t* of *o* is *p* if and only if, for some *D* applying to the relevant situation at *t*, it is a law that the probability of a *D* being an *O* is *p*, and *D* is maximal in the sense that, for any further *E* applying at *t*, the probability of a *D&E* being *O* is *p* as well. (Note how the maximality of such a *D* explains why *D* will exhaust all the *relevant* features of a particular situation; for if *D* is maximal, then any further *E*s will not alter the relative probability of *o*.)

Note that the relative principle concurs with the single-case principle in those cases where agents have information that fixes the single-case probabilities. This follows from the fact that single-case probabilities are special cases of relative probabilities. Consider agents who know all the relevant features of the particular decision situation. The relative probabilities of outcomes fixed by these agents' knowledge of the set-up will coincide with their single-case probabilities, and so these agents will conform to the single-case principle by conforming to the relative principle. The above examples make it clear, however, that this is not the only kind of case where the relative principle specifies the correct degrees of belief on which to act. Often agents' relative probabilities will not be single-case probabilities, yet agents should act on these probabilities in just the same way.

The relative principle accommodates any number of sensible choices that the single-case principle excludes. When informed people make real-life choices (whether to stop smoking, say, or to take some medication, or not to draw to an inside straight), they characteristically act in line with the relative principle but not the single-case principle. However detailed their knowledge of their specific situations, there are always further unknown features (about their metabolism, say, or the next card in the pack) that mean their relative probabilities are not single-case probabilities. But this is no reason to condemn their actions as prudentially incorrect.

There is more to be said about the choice between the relative and single-case principles, and we shall return to this in sections x-xiii below. But first we would like to clarify some other points.

VIII. THE RANGE OF THE RELATIVE PRINCIPLE

Some readers might be tempted to raise the following objection. It seems implausible to suppose that there is a probabilistic law of the form 'the probability of a *D* being an *O* is *p*' for *every* property *D* that the agent might know the set-up in question to have. But where there is no relevant probabilistic law to be had, the relative principle remains silent about how the agent's degree of belief in the outcome is objectively constrained.

Suppose, for instance, that you are presented with a black box that you are told makes coins and asked for your degree of belief that the next coin it produces will land heads when tossed. You know nothing further about the box. You just know it is a black box that manufactures coins. We take it that there is no true probabilistic law stating that the probability of heads on a coin from any coin-manufacturing black box is *p*. The category of coin-manufacturing black boxes is too open-ended and heterogeneous to sustain any serious general patterns. So, given our explanation of relative probabilities in the last section, there is no objective probability of heads relative to your knowledge of the set-up, which means that the relative principle specifies no prudentially correct degree of belief in this case.

Does this not show that the relative principle does not cover all decision situations, and hence that it does not say everything there is to be said about how degrees of belief in outcomes are objectively constrained? We say "no." We are happy simply to concede that not all degrees of belief are so constrained. If the features of the set-up the agent knows about do not constitute the antecedent of any probabilistic law about the outcome in question, then there is no objective constraint on the agent's degree of belief in that outcome. In the above example, there is no prudentially correct degree of belief in heads on the next coin from the black box.

IX PHILOSOPHICAL INTERPRETATIONS OF PROBABILITY

Most relative probabilities are not single-case probabilities. Does this mean that the notion of relative probability presupposes the frequency interpretation of probability? If so, the notion of relative probability is in trouble, for the frequency interpretation of probability is no good.[8] We do not think this is a good reason for rejecting relative probabilities. The topics discussed here, and, in particular, the notions of relative probability and nonmaximal probabilistic law, are independent of disputes among alternative philosophical interpretations of probability. It is true that, if you are interested in philosophical interpretations of probability, then the frequency interpretation, which identifies the probability of an *A*'s being a *B* with the limiting frequency of *A*s in the sequence of *B*s, offers the most immediate explanation of relative probabilities and nonmaximal probabilistic laws. It is also true that the currently most popular alternatives to the frequency interpretation, namely, *chance* or *propensity* theories that take the notion of single-case probability as primitive,[9] have to do some work to account for relative probabilities and nonmaximal probabilistic laws. But we certainly do not think that the notion of relative probability stands or falls (and so falls) with the frequency interpretation.

---

[8] For a quick account of the central flaw in the frequency interpretation of probability, see Papineau, "Probabilities and the Many-Minds Interpretation of Quantum Mechanics," *Analysis*, LV (1995): 239-46.

On the contrary, it seems to us a boundary condition on *any* satisfactory interpretation of probability that it make good sense of relative probabilities and nonmaximal probabilistic laws. Note in this connection that standard statistical tests are indifferent to the maximality or otherwise of the probabilistic laws under investigation. There are well-established empirical methods for ascertaining the probability that a certain kind of machine part, say, will fail in circumstances of some kind. It does not matter to these methods whether or not there are further probabilistically relevant differences between parts and circumstances of those kinds.

Perhaps this is a good reason for preferring those versions of chance or propensity theories which, in effect, take our notion of relative probability as primitive rather than single-case probability.[10] But we shall not pursue this topic further here. We shall simply take the notion of relative probability and probabilistic law as given, and leave the philosophical interpretations of probability to look after themselves.

X. IN DEFENSE OF THE RELATIVE PRINCIPLE

We claim that the relative principle is the fundamental prudential truth about the correct degrees of belief for agents to act on. It may strike some readers as odd to accord the relative principle this authority. After all, many agents will have relative probabilities that differ from the single-case probabilities, and, as a result, the relative principle may specify that they should choose actions other than those which maximize the single-case expected utility.[11]

---

[9] For example, Lewis; R. Giere, "Objective Single-case Probabilities and the Foundations of Statistics," in P. Suppes et alia, eds., *Logic, Methodology and Philosophy of Science* (Amsterdam: North-Holland, 1973); Fetzer, "Reichenbach, Reference Classes, and Single-case Probabilities," *Synthese*, II (1977). 185-217

[10] Cf. I. Hacking, *Logic of Statistical Inference* (New York: Cambridge, 1965); Levi, *Gambling with Truth* (New York: Knopf, 1967); D. Gillies, "Popper's Contribution to the Philosophy of Probability," in A. O'Hear, ed., *Karl Popper* (New York: Cambridge, 1995).

For example, imagine that you accept Alma's £2 to your £1 that 'heads' will be printed out by a device from the machine that makes devices *A* and *B*. You do so on the grounds that your probability for 'heads' relative to your knowledge of the situation is 0.5. But, in fact, this is a device *B*, for which the single-case probability of 'heads' is already 0.2. So declining Alma's bet would have a higher *single-case* expected utility than accepting it (0p instead of -40p).

Here, the relative principle seems to advise the choice that is objectively inferior. Why then do we say it is superior to the single-case principle? On the face of things, it looks as if the relative principle will only give the right advice in those special cases where it coincides with the single-case principle.

Despite this, we maintain that the relative principle says all we need to say about correct degrees of belief. Even in cases where your relative probabilities do not coincide with single-case probabilities, it is objectively correct for you to set your degrees of belief equal to the relevant relative probabilities, and to act accordingly.

Our attention can be distracted here by a point we shall later discuss at some length. There is a sense in which it would be prudentially better if you could *find out* about the single-case probabilities before you choose between the alternative actions on offer. But this does not show that there is anything incorrect about setting your degrees of belief equal to merely relative probabilities when you *have not* found out those single-case probabilities.

Indeed, as we shall see, the explanation of *why* it is prudentially better to find out about the single-case probabilities, if you can, is itself a consequence of the correctness of making choices on the basis of relative probabilities. We shall argue (in sections XIX-XX, following Frank Ramsey and I. J. Good) that the best understanding of why it is *better* to act on the single-case probabilities, if you can, is simply that it is correct, in the sense of the relative principle, to find out the single-case probabilities before acting, if you have the choice. It is certainly true that it is prudentially better to find out the single-case probabilities before deciding, if you can. But this truth, far from challenging the relative principle's notion of correctness, depends on it.

---

[11] When we speak of "an agent's relative probability" for some outcome, this should be understood as shorthand for the relative probability of that outcome in relation to *D*, where *D* is the agent's knowledge of the set-up in question.

XI. A USEFUL COMPARISON

We shall return to these complexities in due course. At this stage, however, some readers may still need persuading that there is *anything* worthwhile about the relative principle when it diverges from the single-case principle. How can it be correct to act on relative probabilities when they advise actions different from those which would be advised by the single-case probabilities, as in the case where you accept Alma's bet not knowing the device in front of you is, in fact, a *B*?

The following comparison may be helpful. First, suppose that, as above, you are offered Alma's bet on a device you know to be from the machine that makes *A*s and *B*s, but where you do not know which sort your device is. Now, compare yourself with Dorinda, say, who is offered the same bet, but in connection with a device that she knows will print out 'heads' with a quantum-mechanical single-case probability of 0.5.

You and Dorinda will both accept Alma's bet, for you and Dorinda will both act on a probability of 0.5 for 'heads'. But where Dorinda is acting on a single-case probability, you are acting on a relative probability instead. Now, would you rather be offered Dorinda's bet than your bet? No. People who make your bet will make money just as fast as people who make Dorinda's bet. Even though Dorinda's bet is in line with the single-case probabilities, and yours is not, yours is objectively just as good a bet. Both of you are quite rightly wagering £1 because your expected repayment is £1.50. (How *can* yours be a good bet, given that it is possible you are betting on a device *B*, which would mean your *single-case* expected repayment is only 60p? But note that we could equally well ask: How *can* Dorinda's bet can be a good bet, given that it is possible her device will in fact print out 'tails', which would mean her repayment is going to be 0p?)

We take this to show that it is objectively worthwhile to conform to the relative principle, even in cases where it could lead you to choose an action that will not, in fact, maximize the single-case expected utility. When you accept Alma's bet, knowing only the relative probabilities but not whether you are betting on a *B* or an *A*, you are making just as good a choice as Dorinda, who knows the single-case probabilities but not what the result will actually be. So, if Dorinda is making a good choice, from a prudential point of view, then so are you.

XII. "A GOOD BET"

It will helpful to expand briefly at this point on a notion that we assumed in the last section—the notion of an objectively "good bet," a bet that is worth accepting instead of declining. This notion encapsulates the idea of the objectively correct choice which we are trying to explicate.

What makes a bet good? An initial temptation is to say that what makes a bet good is that it *will* make you money (a temptation to which we succumbed in the last paragraph, in the interests of dramatic emphasis, when we said that "People who make your bet will make money just as fast as people who make Dorinda's bet"). But, of course, we cannot strictly say this, since even a good bet may turn out unluckily (your device may print out 'tails'). Another temptation is to say that a good bet will make money *in the long run*. But this is little better, since any finitely long run of bets can still turn out unluckily.[12]

Once we resist these temptations, we realize that the only thing that can be said about an objectively good bet is that it is objectively *probable* it will make you money; more generally, that the objectively right choice is that which offers the maximum *objective expected utility* among the alternatives open to you. It may be surprising that there is nothing more to be said, but it is a conclusion that a long tradition of philosophers addressing this question have been unable to avoid.[13]

What is even more surprising, perhaps, is that it does not matter to the objective correctness of a choice whether the objective probabilities that make it correct are single-case or not. All that matters is that they are the objective probabilities relative to your knowledge of the decision situation. The comparison in the last section makes this point clear. It does not matter to the goodness of the bet offered by Alma whether the single-case probability of heads is currently different from 0.5. As long as you know the decision situation as one in which the relative probability of heads is 0.5, it will be right for you to accept the bet; and you will be just as likely to make money, in the objective sense of probability that matters to choice, as someone who accepts a similarly structured bet on the basis of knowledge that the single-case probability is 0.5.[14]

---

[12] If you believe in the frequency theory of probability, you can argue that a good bet will definitely make money in the *infinite* long run. But, even for frequentists, it is not clear why this should *make* a bet good, given that none of us is going to bet forever.

[13] Cf. C.S. Peirce, "The Doctrine of Chance," in M.R. Cohen, ed., *Chance, Love, and Logic* (New York: Harcourt, 1924); and Hilary Putnam, *The Many Faces of Realism* (La Salle: Open Court, 1987).

XIII. DECISION SITUATIONS

One way of conveying our picture of prudential rationality would be to say that the practical predicament facing an agent cannot be separated from the agent's knowledge of his surroundings. Let us use the admittedly inelegant term *decision situation* for the context of a choice. It is tempting to think of such a decision situation as consisting solely of the objective set-up that will generate the outcomes of interest—such as, for example, some physical device that will print 'heads' or 'tails' with certain probabilities.

If we think of decision situations in this way, then it is natural to conclude that, if there are correct degrees of belief, they must be those which correspond to the current single-case probabilities. For there is no way of identifying a unique probability for an outcome, given only the set-up in all its particularity, except as the probability fixed by all current features of the set-up—that is, the current single-case probabilities.

But suppose we think of the decision situation facing an agent as consisting of a set-up *plus* the properties the agent knows the set-up to have. Now, there is another way of identifying a unique probability for an outcome—namely, the outcome's probability relative to all those properties which constitute the decision situation.

Note that on this view it is still an objective feature of the world which makes an agent's degree of belief correct. For it is a perfectly objective fact that the probability of outcome *o* relative to some property *D* of the set-up is *p*. Many such facts obtain quite independently of what agents know, or even of whether there are agents at all. Of course, *which* such fact makes a particular agent's degree of belief correct will depend on that agent's decision situation. But it is scarcely surprising that which objective probabilities matter to decisions should be different for different decision situations.

---

[14] Is there not a danger, when betting on merely relative probabilities, that your opponent may know the single-case probabilities, and so have the advantage of you? (Alma may have a secret way of telling if the device is an *A*.) But note that in such betting situations you will have extra information apart from knowing about the structure of the betting device—you will also know, for example, that Alma is offering 2-1 against heads. So the recommendation of the relative principle in such a case depends crucially on the objective probability of heads *relative to the fact that someone is offering these odds*. If this is still 0.5, then your bet is fine. But if it is significantly less than 0.5, as would be the case if people like Alma have a way of fleecing suckers, then you should avoid such bets.

We would all do well to remember the advice given by Sky Masterson's father in Damon Runyon's "The Idyll of Miss Sarah Brown" (in *Runyon a la Carte* (London: Constable, 1946)): "Son, no matter how far you travel, or how smart you get, always remember this: Some day, somewhere, a guy is going to come to you and show you a nice brand-new deck of cards on which the seal is never broken, and this guy is going to offer to bet you that the jack of spades will jump out of this deck and squirt cider in your ear. But, son, do not bet him, for as sure as you do you are going to get an ear full of cider." In what follows, however, we shall simplify the exposition by continuing to assume, contra Mr. Masterson, Sr., and common sense, that the existence of attractive bets is in general probabilistically irrelevant to the relevant outcomes.

XIV. THE PRINCIPAL PRINCIPLE

It is an implication of Lewis's "principal principle" (*op. cit.*) that, if you know the single-case probability of some outcome and nothing else relevant (nothing "inadmissible" in Lewis's terminology), then you ought to set your degree of belief equal to that single-case probability. It will be illuminating to compare the position we are defending here with Lewis's.

One immediate point is that Lewis seems as much concerned with issues of evidential rationality as with prudential rationality. His principal principle is part of a more general Bayesian account of the degrees of belief required of rational thinkers who have such-and-such evidence, and bears on questions of the prudential correctness of degrees of belief, if at all, only derivatively.

Even so, Lewis holds that connection forged by the principal principle between degrees of belief and single-case probabilities exhausts the ways in which degrees of belief are rationally constrained by knowledge of objective probabilities. So it is natural to ask what Lewis would say about agents who know relative probabilities but not single-case probabilities. Take the case of the machine that makes devices *A* and *B*. As before, you know that the device in front of you comes from this machine but not whether it is *A* or *B*. So you know that the probability of 'heads' relative to your information is 0.5, but you are ignorant of the single-case probability of this outcome. Lewis's principal principle would seem to have no grip here. Yet, surely, evidential rationality, just as much as prudential rationality, requires you to have a degree of belief of 0.5 in heads.[15]

Lewis's response to this challenge[16] hinges on the fact that the principal principle, unlike the single-case principle, does not refer specifically to *current* single-case probabilities. Rather, it says that, if you know the value of the single-case probability for any time *t*, and nothing else relevant, then you ought to have a matching degree of belief, even at times other than *t*. In the case at hand, Lewis would argue that you know the single-case probability of heads was once 0.5 (before the machine opted between *A* and *B*), and you know nothing else relevant (since your later information tells you nothing about which device the machine made). So the principal principle requires you to set your degree of belief equal to the relative probability of 0.5 after all.

---

[15] Cf. Levi, Review of R.C. Jeffrey, ed., *Studies in Inductive Logic and Probability, Volume II* [*op. cit.*], in *Philosophical Review*, XCII (1983): 120-21; and Kyburg, "Principle Investigation," this JOURNAL, LXXVIII, 12 (December 1981): 772-78.

[16] See Lewis, "Postscripts to 'A Subjectivist's Guide to Objective Chance'," in his *Philosophical Papers, Volume II* (New York: Oxford, 1986), pp. 114-32.

XV. A MODIFIED SINGLE-CASE PRINCIPLE?

The apparent success of Lewis's response might suggest that we were perhaps too quick to replace the single-case principle by the relative principle. Perhaps we should have tried:

> (5) The *modified single-case principle*: the correct subjective probabilities for an agent to attach to outcomes $\{o_i\}$ are equal to the single-case probabilities of the $\{o_i\}$ at that earlier time when nothing probabilistically relevant to *o* was yet determined, apart from those features which the agent now knows about.

This would not require that the correct degrees of belief are the *current* single-case probabilities. Indeed, this principle would agree with the relative principle, to the extent that it specifies that the correct degrees of belief are a function, inter alia, of which facts the agent now knows about. But it would at least retain the original idea of equating prudentially correct degrees of belief with single-case rather than relative probabilities.

This modified single-case principle will not do, however. Consider once more the case where you do not know whether the machine has given you device *A* or *B*. But now suppose that you have observed your device print out 'heads' as its first two results. What is the correct degree of belief in 'heads' next time? A simple calculation using Bayes's theorem will show that the probability of 'heads' next time on a device that has already produced two 'heads' is 13/17. This probabilistic law is a consequence of the other probabilistic laws already specified for this example. So the relative principle specifies, quite rightly, that the correct degree of belief for 'heads' on the next toss is around 13/17. Somebody who uses this figure to decide which bets to accept will make the correct choices.

But note that 13/17 is not, *nor ever has been*, the single-case probability of 'heads' next time. The single-case probability of heads on any toss was 0.5 before the device was made, and is now either 0.2 or 0.8. So even the modified single-case principle cannot cover this case. The only fully general principle governing prudentially correct degrees of belief is that they should equal the relative probabilities.

We should stress that this example does not refute Lewis's view. For Lewis is not committed to the modified single-case principle as the only principle governing reasonable degrees of belief. Rather, Lewis's full theory of reasonable degrees of belief contains two principles that together accommodate cases that escape the modified single-case principle: these are (a) the principal principle by which he links reasonable degrees of belief to single-case probabilities, and (b) a principle of Bayesian conditionalization for updating reasonable degrees of belief. So he can deal with the present example by saying that rational agents will first set their degrees of belief for *A* and *B* equal to 0.5, in line with the principal principle; and then they will update these degrees of belief by Bayesian conditionalization to get the right answer after they observe that their device has produced two 'heads' already.

Still, apart from demonstrating that prudentially correct degrees of belief do not always match single-case probabilities (and thus that the modified single-case principle on its own is inadequate), our example also shows that there is a measure of theoretical economy to be gained by upholding the relative principle rather than Lewis's principal-principle-plus-Bayesian-conditionalization. The advantage is that we do not need to adopt conditionalization as a separate principle, at least in cases structured by objective probabilistic laws. For it can be show that, in such cases, conditionalization falls out of the relative principle, in the way illustrated by our example: if you set your new degree of belief equal to the relative probability of getting the outcome in situations satisfying your initial-knowledge-plus-new-information, then you will automatically be conditionalizing. By contrast, note that conformity to conditionalization does not follow from the principal principle, but must be added as an independent requirement on rationality, since updated degrees of belief often will not match single-case probabilities, as in our example. It seems to us preferable to avoid postulating Bayesian conditionalization as an extra requirement, given the well-known difficulties of providing this principle with a satisfactory rationale.

XVI. LEWIS AND DETERMINISM

We would like to offer a different and more direct reason for preferring our relative principle to Lewis's package of principal-principle-plus-Bayesian-conditionalization. Consider Elspeth, a convinced determinist, who is asked to bet on a certain coin coming down heads. Elspeth has performed a large number of trials on coins of this manufacture and has overwhelming statistical evidence for a probabilistic law that says that the probability of heads with coins just like this one is 0.9. Suppose also that determinism really is true. What ought Elspeth's degree of belief in the coin's landing heads on the next throw to be?[17]

Since Elspeth thinks that heads has a single-case chance of either 0 or 1, the full version of Lewis's principal principle implies that her degree of belief in heads should be the average of these two chances weighted by her degree of belief in each, which will, of course, equal her degree of belief in the latter chance. But what ought this degree of belief to be? To us it seems clear that it should equal 0.9. But there is nothing in Lewis's principal principle, nor in his principle of Bayesian conditionalization, to say why. The relative principle, on the other hand, can deal with the case: the objective probability of heads relative to Elspeth's knowledge of the situation is 0.9, so 0.9 is the correct degree of belief for Elspeth to have.

There are various responses open to Lewis. An initial possibility would simply be to deny that there would be any probabilistic laws if determinism were true; but this seems to us a thesis that needs some independent argument in its favor. An alternative is to appeal to the notion of "counterfeit chance"—a degree of credence that is resilient in the sense that it would not be undermined by further "feasible" investigation—as a surrogate for the notion of a probabilistic law under determinism.[18]

Counterfeit chances have an indeterminacy, however, that makes Lewis suspicious of their credentials as constraints on degrees of belief. He says:

> Counterfeit chances will be relative to partitions; and relative, therefore, to standards of feasibility and naturalness; and therefore indeterminate unless the standards are somehow settled, or at least settled well enough that all remaining candidates for the partition will yield the same answers. Counterfeit chances are therefore not the sort of thing we would want to find in our fundamental physical theories, or even in our theories of radioactive decay, and the like. But they will do to serve the conversational needs of determinist gamblers.[19]

Put in our terms, Lewis is here appealing to the point that probabilistic laws will often imply different probabilities for the same outcome, depending on what nonmaximal descriptions of the set-up we employ. Lewis seems to be suggesting that such probabilities cannot therefore place objective constraints on agents' degrees of belief. But there is no good basis for this conclusion. It is only if you assume that such constraints must always be supplied by single-case probabilities that you will find their relativity to descriptions worrying. By contrast, this relativity is unproblematic from our point of view, since our tie between degrees of belief and objective probabilities is forged by the relative principle, not the principal principle, and this explicitly specifies that correct degrees of belief are relative to agents' knowledge of the set-up.

---

[17] Cf. Levi, Review of Jeffrey, and "Chance," *Philosophical Topics*, XVIII (1990): 117-49.

[18] Cf. Bryan Skyrms, *Causal Necessity* (New Haven: Yale, 1980); and Lewis, "Postscripts to 'A Subjectivist's Guide to Objective Chance'."

[19] Lewis, "Postscripts to 'A Subjectivist's Guide to Objective Chance'," p. 121.

Lewis gives no good reason to regard nonmaximal probabilistic laws as second-class citizens. We may not want to find such laws in our fundamental physical theories, but that does not mean that they are *only* capable of serving the conversational needs of determinist gamblers. On the contrary, the part that nonmaximal probabilistic laws play in the relative principle makes them fundamentally important in understanding how degrees of belief in outcomes are objectively constrained. To deny this is to deny that there is any objective feature of the world which makes it correct for Elspeth to attach a 0.9 degree of belief to heads.

XVII. SALMON'S VIEW

Salmon argues forcefully that subjective degrees of belief ought to match objective probabilities, in addition to satisfying internal constraints of coherence and conditionalization (*op. cit.*). Although Salmon thinks of objective probabilities as long-run frequencies, he is sensitive to the distinction between single-case and relative objective probabilities: for him, this is the distinction between frequencies in objectively homogeneous reference classes (genuine *chances*) and frequencies in merely epistemically homogeneous reference classes.

When Salmon addresses this distinction (*op. cit.*, pp. 23-24), he argues that the fundamental external constraint on degrees of belief is that they should match the chances (the single-case probabilities); that is, his basic commitment is to the single-case principle. He does recognize, however, that we need some account of the worth of choices informed by merely relative probabilities. His response is to view degrees of belief that match merely relative probabilities as the *best available estimates* of single-case probabilities. In effect, he views the relative principle as a matter of trying to conform to the single-case principle under conditions of limited information.

This does not work. Consider again the machine that makes *A*s and *B*s. You have a device from this machine, but you do not know which kind it is. Your relative probability for 'heads' is 0.5. This does not seem a good estimate of the single-case probability—after all, the single-case probability is certainly either 0.2 or 0.8. At best, Salmon could argue that the relative probability is a weighted average of the possible single-case probabilities. But then the weighting factors (0.5 for the 0.2 single-case probability, and 0.5 for the 0.8) are relative probabilities once more, and we are left once more with an unexplained appeal to the worth of matching degrees of belief to relative probabilities. Whichever way we turn it, Salmon does not give us a way of explaining the relative principle in terms of the single-case principle.

XVIII. MELLOR'S VIEW

In *The Matter of Chance*, Mellor argues that the correct degrees of belief are those which match the chances, and to that extent is in prima facie agreement with our single-case principle (*op. cit.*). He then considers an example that is effectively equivalent to our machine that makes *A*s and *B*s. In connection with this example, he appeals to another thesis defended in his book, namely, that chances are functions not just of specific outcomes ('heads' next time) but also of the *kind of trial* generating the outcome. So, if the trial is (a) *the machine makes a device that displays a result*, then the chance of 'heads' is 0.5; but if it is (b) *this particular device (that came from the machine) displays a result*, then the chance of 'heads' is either 0.2 or 0.8.

So Mellor makes chances relative to the kind of trial. What determines the kind of trial? If this depends on the agent's *knowledge* of the set-up, in such a way that the trial is of kind (a) whenever the agent does not know whether the device is an *A* or a *B*, and is of kind (b) whenever the agent does know which it is, then Mellor has effectively switched to the relative principle, for his "chances" are equal to probabilities relative to the features of the set-up the agent knows about.

Mellor specifies, however, that his "chances" are not relative to knowledge, but rather to what "counts as repeating the bet" at hand: if the bet is on the *device-generating machine*, so to speak, the "chance" is 0.5, while if it is on *this device*, the "chance" is 0.2 or 0.8. Mellor's resulting position is certainly different from the relative principle, since repetitions of a bet can presumably be on *this device* rather than the machine, even though the agent does not know whether this device is an *A* or *B*. But for just this reason, Mellor's position is unsatisfactory, since he is left without any explanation of why the prudentially correct degree of belief for this kind of agent is 0.5.

In his more recent "Chances and Degrees of Belief," Mellor returns to this issue, and argues that such an agent should act on a 0.5

degree of belief after all, since of the two "chances" that attach to the outcome 'heads' (0.5 relative to trial (a); 0.2 or 0.8 relative to trial (b)), this is the only one that the agent knows (*op. cit.*). We agree that this is the right answer, but would observe that Mellor is only able to give it because he has switched to the relative principle. In our terms, he now holds that agents should act on the probabilities relative to the features of set-up that they know about, that is, that their degrees of belief should match their relative probabilities.

XIX. MORE INFORMATION IS BETTER

We have argued that, if your degrees of belief match your relative probabilities, it is irrelevant to the goodness of your consequent choice whether or not these are also single-case probabilities. Degrees of belief that correspond to the relative probabilities on one gambling device can be just as correct, from a prudential point of view, as numerically identical degrees of belief that match single-case probabilities on a different device.

As we admitted earlier, however, there is a different sense in which it *is* better to act on single-case probabilities, if you can find out about them. To pin down this sense, it will be helpful to consider, not two different gambling devices, but a single device, about which you might have more or less information. Imagine, for example, that you are facing a device from the machine that makes *A*s and *B*s. You are able *either* to act on your present 0.5 relative probability for heads on a device from this machine, *or* to find out whether your device is an *A* or a *B*, and so act on the single-case probability of heads. In this kind of comparison, where the single-case probability involves *more* information about the *same* device, we agree that it is prudentially better to find out the single-case probability before deciding.

We deny, however, that this point calls for any modification of, or addition to, the relative principle. That it is prudentially better to find out the single-case probabilities if you can, before committing yourself to action, is itself a consequence of the relative principle.

What is more, this corollary of the relative principle is simply a special case of a more general point, namely, that it is always prudentially better to act on probabilities relative to more rather than less information about a given set-up, if you can. Not only are single-case probabilities better than relative probabilities, but relative probabilities that are relative to greater amounts of information are better than relative probabilities that are relative to less. Indeed, the same point applies within the category of single-case probabilities. As long as you can keep your options open, it is prudentially better to wait and see how the single-case probabilities evolve, instead of deciding straight off knowing only their initial values. Best of all, of course, is to wait until you know the outcome itself. If you can delay until Alma has removed her hand, then you can accept her bet whenever the coin shows heads, and decline whenever it shows tails.

XX. THE RAMSEY-GOOD RESULT

Let us first offer an example to illustrate why it is always prudentially better to act on probabilities relative to more rather than less information about a given set-up, if you can. We shall then sketch a general demonstration.

Imagine you are presented with three sealed boxes. You know that a ball has been placed in one, and that the other two are empty, and that some chance mechanism has given each box the same 1/3 single-case probability of getting the ball. Felicity offers you her £3 to your £2, with you winning the total stake if the ball is in box 1. You can either accept or decline the bet now (option *G*, for "go"), or you can look in box 3 and then accept or decline the same bet (option *W*, for "wait-and-see").

If you take option *G* and choose now, you will decline Felicity's bet (since you would be putting up £2 for an expected return of £1.67). So, since you will decline, the expected gain of option *G* is £0.

If you take option *W*, then 1/3 of the time you will find the ball in box 3, and so decline the bet, since you would then be sure to lose it. But 2/3 of the time you will not find the ball in box 3, and in these cases you will accept the bet, since the expected return on your £2 will now be £2.50. Averaging over these two possibilities, weighted by their respective probabilities, you can expect to gain £0.33 if you take option *W*.

So option *W* is better than option *G*, since its expected value is £0.33 rather than £0. The reason is that option *W* gives you the opportunity of switching to accepting the bet in just the case where your extra information (no ball in box 3) indicates that this is worthwhile.

This example illustrates the general result. Suppose you are faced with alternative actions $\{A_i\}$, and have a metachoice between deciding now (option *G*), or finding out whether *C* before acting (option *W*). If this extra information can affect which A_i you choose, and if the cost of acquiring this information is negligible, then the expected gain from option *W* is always greater than the expected gain from option *G*.

We shall sketch a proof (following Ramsey and Good[illegible]) for the simple case of two alternative actions A_1 and A_2 (accept or reject

---

[illegible] Ramsey, "Weight or the Value of Knowledge," *British Journal for the Philosophy of Science*, XLI (1990): 1-4; Good, "On the Principle of Total Evidence," *British Jounal for the Philosophy of Science*, XVII (1967): 319-21 some bet, say). If you take option G, then you will now choose one of these—A_1, say. Let $\mathrm{EU}(A_1/C)$ be the expected utility of this action assuming C. Then we can decompose the expected gain of taking option G and so choosing A_1 as

$$\text{(I)} \quad \mathrm{EU}(G) = \mathrm{EU}(A_1/C)\mathrm{Pr}(C) + \mathrm{EU}(A_1/\text{not-}C)\mathrm{Pr}(\text{not-}C)$$

Now, suppose you take option W instead and find out whether C. The discovery that C might make you choose A_2 rather than A_1. Alternatively, the discovery that not-C might make you do this. (Both discoveries cannot make you switch; otherwise, you would not have preferred A_1 to start with.)

Suppose, without loss of generality, that discovering C makes you switch to A_2. Then

$$\text{(II)} \quad \mathrm{EU}(W) = \mathrm{EU}(A_2/C)\mathrm{Pr}(C) + \mathrm{EU}(A_1/\text{not-}C)\mathrm{Pr}(\text{not-}C).$$

But this must be larger than $\mathrm{EU}(G)$, because if $\mathrm{EU}(A_2/C)$ was not bigger than $\mathrm{EU}(A_1/C)$, you would not have chosen A_2 when you discovered C.

XXI. MORE INFORMATION IS RELATIVELY BETTER

We take this result to explain why it is better to find out the single-case probabilities before acting, when you can. More generally, we take it to explain why it is better to delay acting until you acquire more probabilistically relevant information, whenever you can.

It is noteworthy, if not immediately obvious, that this explanation makes essential appeal to the relative principle rather than to any single-case principle. Waiting-and-seeing is better than going because it has a higher objective expected utility from the point of view of your *relative probabilities prior to this choice*. In equations (I) and (II), $\mathrm{Pr}(C)$ and $\mathrm{Pr}(\text{not-}C)$ are not single-case probabilities, but probabilities relative to your initial information; and the same goes for the conditional probabilities $\mathrm{Pr}(o_i/C)$ and $\mathrm{Pr}(o_i/\text{not-}C)$ implicitly involved in the quantities $\mathrm{EU}(A_i/C)$ and $\mathrm{EU}(A_i/\text{not-}C)$.

We can illustrate the point with the example of the boxes. Here, $\mathrm{Pr}(C) = 1/3$ is the probability of finding the ball in box 3. This is not the single-case probability of that outcome, for that is already either 0 or 1.[21] Rather, it is the probability relative to your initial knowledge of the set-up. Similarly, the conditional probability = 1/2 of the ball being in box 1 (call this outcome o), given it is not in box 3, which is central to the expected utility calculations, is a relative conditional probability. It is calculated by dividing Pr(o & not-C) = 1/3 by Pr(not-C) = 2/3, where these are again all probabilities relative to your initial information. (The *single-case* conditional probability of o given not-C—that is, Sing-CPr(o & not-C)/Sing-CPr(not-C)—will already either be 0 or 1 or undefined—depending on whether the ball is in fact in box 2, box 1, or box 3, respectively.)

---

[21] Was not 1/3 *once* the single case probability of C? Well, then consider an example where your relative probabilities of 1/3 for each box is the result of conditionalizing on previous results, as in the kind of case discussed in section XV above.

In effect, the Ramsey-Good result reduces the *superiority* of decisions based on single-case probabilities, and better informed probabilities generally, to the *correctness* of waiting-and-seeing instead of going, whenever you have this choice. But, for this result to work, correctness must be understood in the sense of the relative principle, not the single-case principle.[22]

XXII. FAILING TO MAXIMIZE SINGLE-CASE EXPECTED UTILITY

We can highlight the dependence of the Ramsey-Good result on the relative principle by noting that the sense in which it is always better for you to acquire additional probabilistically relevant information, if you can, does not guarantee that you will end up choosing the action A, which maximizes the *single-case* expected utility. Consider the above example again, and imagine the ball is in fact in box 2. You look in box 3, see nothing, and so accept Alma's bet, since it is now a good bet. But it was determined from the start that this bet would lose. So, despite the fact that it is better, from the point of view of the probabilities on which you will quite correctly act, to wait-and-see rather than go, the single-case expected utility of waiting-and-seeing in this particular case is -£1, by comparison with the single-case expected utility of £0 for going.

This is simply a reflection at the metalevel of a point made in section X. The action that is specified as correct by the relative principle need not be the action that maximizes the single-case expected utility. Since the Ramsey-Good result explains the superiority of waiting-and-seeing in terms of the relative principle rather than any single-case principle, it is unsurprising that there should be cases where waiting-and-seeing does not maximize single-case expected utility.

---

[22] David Miller complains that the Ramsey-Good result fails to explain the *principle of total evidence* (*Critical Rationalism* (Chicago: Open Court, 1994), chapter 7). Our view is this. We have no explanation of why you should act on the probabilities relative to *all* your information. That is the relative principle, and we take this to be a primitive truth about prudential rationality. But this principle then explains, via the Ramsey-Good result, why you should arrange to act on probabilities relative to *more* information, if you can.

XXIII. FINDING OUT THE SINGLE-CASE PROBABILITIES

Of course, in the special case where acquiring more information tells you about the current single-case probabilities, instead of just giving you better informed relative probabilities, then the A, chosen after waiting-and-seeing will indeed maximize the single-case expected utility, since you will then be acting on the single-case probabilities themselves. (Imagine that you can look in *two* boxes, or find out whether your device is A or B, before deciding on Alma's bet.) This may make some readers feel that in cases where you find out the single-case probabilities, at least, the rationale for waiting-and-seeing is independent of the relative principle. Surely, here the justification for waiting-and-seeing is simply that this will enable you to act on the single-case probabilities.

This thought simply takes it for granted, however, that it is better to act on the single-case probabilities if you can, without offering any explanation for this. We think it better to explain things, if you have the materials to do so, rather than take them as primitive. The Ramsey-Good result explains why it is better to act on the single-case probabilities, rather than relative probabilities, when you have the choice. But it does so by assuming the correctness of acting on your initial relative probabilities rather than on your final single-case probabilities.

XXIV. CONCLUSION

It is natural to think that the prudentially correct action is always the action that maximizes single-case expected utility. This, in turn, makes it natural to suppose that the single-case principle specifies the basic notion of prudential correctness, and that the relative principle is at best some kind of derivative poor relation.

Our analysis indicates, however, that this has things the wrong way round. To see this, suppose you postulate the single-case principle as a basic principle governing prudential correctness. You will then need to say something about agents who act on relative probabilities. Since relative probabilities need not be (nor need ever have been) equal to single-case probabilities, you will need separately to introduce another notion of prudential correctness, as in the relative principle. The question will then arise of why it is better to be correct in the first single-case sense than in the second relative sense. You could take this to be a further primitive fact. But it would make more sense to explain it, via the Ramsey-Good result, as itself a consequence of the relative notion of correctness, since you will in any case need this explanation for the more general point that it is better to act on more probabilistically relevant information than less.

By this stage, however, the single-case principle will have ceased to do any work. For if we simply begin with the relative principle, which we have to assume anyway, we do not need separately to postulate a notion of single-case correctness, since agents who act on single-case probabilities are per se acting on relative probabilities. The relative principle itself will explain why it is preferable to act on single-case probabilities and, in general, on probabilities relative to more information, when you can.

It may seem unnatural to regard the relative principle as more basic than the single-case principle. But this position has the support of theoretical simplicity.

HELEN BEEBEE

University College/London

DAVID PAPINEAU

King's College/London

A SHOOTING-ROOM VIEW OF DOOMSDAY

Recently, H. B. Nielsen, John Leslie, and J. Richard Gott, III[1] have published similar arguments purporting to show that the time left until the end of the human race may be shorter than we generally suppose. In its simplest form, this reasoning can be summarized as follows. We may think of ourselves as randomly selected from among all who ever live. Our randomness makes it unlikely that we are among the earliest humans. A calculation demonstrates that, if we are not among the earliest humans, then the human race has less time left than is usually allotted to it.

The treatments differ mainly in emphasis. Gott's article has more mathematical trappings and brims with speculation; he also considers questions other than that of human survival. Leslie presents a dazzling array of illustrations and analogies, evidently as an aid to negotiating this treacherous terrain. He is far and away the most prolific on this topic, having written a dozen articles as well as a book-length treatment of issues relating to the end of the world.[2] Following Leslie, we call this kind of reasoning the *doomsday argument*.

The ease with which the doomsday argument extracts substantial information about the remote future from a quite modest investment of current fact has provoked suspicion in most quarters; however, critiques that have not been snidely dismissive have tended to be as mystifying as the doomsday argument itself. My purpose here is to show that the fallacy inherent in the argument can be discerned by focusing on correct application of Bayes's theorem. It facilitates the analysis that doomsday reasoning closely parallels a certain fallacious train of thought associated with a probability-theory puzzle called the *shooting room*.

---

[1] Nielsen, "Random Dynamics and Relations between the Number of Fermion Generations and the Fine Structure Constants," *Acta Physica Polonica* (B) XX, 5 (1989): 427-68. Leslie, "Is the End of the World Nigh?" *The Philosophical Quarterly*, XL, 158 (1990): 65-72; "Time and the Anthropic Principle," *Mind*, CI, 403 (1992)· 521-40; "Doom and Probabilities," *Mind*, CII, 407 (1993): 489-91, *The End of the World* (New York: Routledge, 1996). Gott, "Implications of the Copernican Principle for Our Future Prospects," *Nature*, CCCLXIII (1993): 315-19. For a wholly different approach to the question of assigning probabilities to doomlike events, see Martin H. Krieger, "Could the Probability of Doom Be Zero or One?" this JOURNAL, XCII, 7 (July 1995): 382-87.

[2] For references, see Leslie, *The End of the World*.

0022-362X/97/9405/244-59

Leslie reviews a variety of objections to the doomsday argument, but I believe these distract from the core issue.[3] Many concern some intrinsic unknowability of the future; for example, doomsday scenarios inhabit a limbo of unrealized possibilities, thereby evading all current attempts at divination. The difficulty in mounting an attack on the doomsday argument solely on the basis of some alleged intractability of future events—that they are indeterminate, unactualized, multipotential, and so on—is that success would necessitate discarding the whole project of making future projections based on current information. When a demographer makes a projection of the world population in the year A.D. 2100, it is quite reasonable to challenge the data, the model, or the computation. But does it make sense to maintain the estimate is invalid because physics may be indeterministic or because these projections are merely of possible people, whereas in the year A.D. 2100 the world will be populated with real people? It would be calamitous if we had to refrain from employing probabilistic or statistical reasoning regarding future populations until elementary particle physicists came to a decision as to whether the world is indeterministic or until philosophers settled the numerosity of possible futures or the nature of unborn identity. If the doomsday argument makes questionable projections, it is on account of flaws in the procedure, not because of some generic defect in the idea of estimating unborn populations. Instead of wrangling over the metaphysics of measuring a fate that has not yet been sealed, I choose to examine the validity of the probabilistic reasoning underlying the doomsday argument.

Leslie's views on the regulating role of determinism in the doomsday and kindred arguments are, I believe, wholly unjustified. According to him, this kind of reasoning works excellently under determinism, but as we slide along the scale of increasing indeterminism, doomsday arguments become progressively undermined until, for the case of radical indeterminism, they may fizzle altogether.[4] If there existed a mode of statistical inference that were valid according to the extent that determinism were true, however, then by repeatedly testing the accuracy of this type of statistical inference, one could gauge the correctness of determinism. Since this conclusion is highly implausible, it is a safe bet that statistical inferences, including those which underlie the doomsday argument, do not hinge on the truth of determinism. (Except for certain quantum experiments, uncertainties from other sources swamp any uncertainty originating in indeterminism; this is why the determinism question is not a burning issue among, say, insurance companies.) Similar counterarguments can be fashioned around some of the other philosophical issues upon which certain critics have claimed the doomsday argument's validity rests—the multiplicity of possible futures, presentist versus eternalist theories of time, and the nature of human identity. It would be a tremendous boon to philosophical investigation if solutions to such problems were connected to the power or accuracy of certain statistical techniques or of applications of Bayes's theorem. Lacking this philosopher's stone, the decision maker must "play" against uncertainty due to ignorance and uncertainty due to indeterminism in the same manner; this applies also to doomsday reasoning. (One might protest that the doomsday situation is unique and cannot be modeled experimentally; however, Leslie is explicit that his novel reasoning about determinism applies to certain games, some of which can be played or simulated. See section III for an example.)

---

[3] As an example, Leslie expends not a little effort countering the somewhat fatuous objection that future humans are not alive to observe anything (*ibid.*, pp. 19-21, 214-18, 246-47, and "Doom and Probabilities," pp. 489-90, in which, unaccountably, this objection is ascribed to me). Surely, it would suffice to point out that the issue is not what unborn humans can or cannot do, but what we can infer about their numerosity from our birth rank.

[4] "Time and the Anthropic Principle," p. 537; *The End of the World*, pp. 188, 233-34.

I. BAYES'S THEOREM

Proponents of the doomsday argument base their reasoning on Bayes's theorem, an uncontroversial, almost trivial proposition of the probability calculus.[7] Use of Bayes's theorem should not be confused with Bayesianism, a controversial viewpoint on the nature of probability associated with subjectivism and the claim that correct statistical inference depends crucially on assessment of prior distributions. The most avid anti-Bayesian has no quarrel with Bayes's theorem proper. Leslie blurs this distinction in a manner that allows him to portray the doomsday argument as an application of a controversial doctrine (Bayesianism) to a straightforward fact of our existence (birth rank) with the suggestion that the argument, though a little dubious, may, along with Bayesianism, turn out to be fundamentally sound. Bayes's theorem is a rigorously demonstrable statement of the probability calculus, however; if the premises of the theorem are fulfilled, the conclusion follows with the force of logic. As with other rigorous results, the sticking point in applications is not whether the theorem is correct, but the extent to which the premises are fulfilled. The doomsday argument should be seen as a straightforward application of an uncontroversial rule (Bayes's) to data produced from a highly questionable assumption (the human randomness assumption, described below).

---

[7] See the first equation in footnote 8; virtually any textbook of probability theory treats this topic—for example, William Feller, *An Introduction to Probability Theory and Its Applications, Volume I* (New York: Wiley, 1968), p. 124.

Leslie seems to believe these matters can be settled through sheer weight of accumulated analogies. Though of possible pedagogic or heuristic value, such analogical reasoning can at best support only the preliminary stages of investigation, whereafter it becomes incumbent to find nonanalogical evidence or to investigate the validity of the analogies themselves. My approach is to concentrate on whether in its doomsday use, the preconditions for valid application of Bayes's theorem have been fulfilled.

I shall begin with a set of alternatives, A_1, A_2,...,A_n... and a piece of information R. For the doomsday argument, we interpret R as knowledge of one's own birth rank[*] and the A_ns to be various doomsday scenarios. To apply Bayes's theorem, we need two sets of probabilities:

(1) *Prior probabilities* for the alternatives A_ns. These probabilities are *logically* prior to R in the sense that they do not take R into account; in applications, the A_ns often refer to events that are also temporally prior to R, but this is not essential.

(2) *Conditional probabilities* connecting R to the A_ns; these tell how likely R is under each of the relevant alternatives.

Bayes's theorem then permits the derivation of a new set of probabilities revealing how likely each of the A_ns are after the information R has been factored in; these new probabilities represent a revision of the prior probabilities in light of the additional information R. If no revision occurs, we speak of a *trivial* application of the theorem; intuitively, this means R has nothing to do with the A_ns.

II. THE FALLACY

In the literature on Bayes's theorem, most attention has been paid to securing the often problematic prior probabilities. By contrast, the role of the conditional probabilities is abundantly clear: if the information R is not probabilistically connected to the alternatives, then learning *R* could not possibly tell us anything new about these alternatives, and applying Bayes's theorem would be fruitless.

---

[*] *Birth rank* refers to one's place in the birth order, numbered consecutively, of all humans who ever live. Of course, we know our own birth rank only very roughly. By assuming knowledge of an exact rank, we simplify the analysis while making a stipulation that is favorable to the argument we are critiquing. If an exact rank is worthless for projecting doomsday, then an approximate one is also.

In the doomsday argument, the conditional probabilities address the question of how doomsday is probabilistically connected to one's own birth rank. The natural, and I believe correct, answer is that there exists no connection whatever. In that case, Bayes's theorem applies trivially and there should be no shift in our assessments of doomsday upon learning our own birth rank. If the doomsdayers draw the opposite conclusion, it is because they possess an additional assumption, namely, the *human randomness* (HR) *assumption*:

> We can validly consider our birth rank as generated by random or equiprobable sampling from the collection of all persons who ever live.

Bayes's theorem permits the derivation of a conditional dependence from the converse dependence, in the presence of the right kind of background information. If our birth rank can tell us via Bayes's theorem something about the likelihood of doomsday, then it has to be because doomsday can tell us something about the likelihood of our birth rank; there is no way around this. The HR assumption provides the needed connection, since in that case the probability one has a given birth rank is inversely proportional to how many ranks exist: the sooner doomsday comes, the fewer the available birth ranks, the more likely a random human has a given rank. It is this which powers the doomsday argument, since under the HR assumption, chancing to have a birth rank as low as we do makes it likely there are relatively few ranks available overall.

Leslie claims that doomsday reasoning "doesn't generate risk estimates *all by itself*. It argues for Bayesian shifts which magnify any risk-estimates that have been reached by other means."[7] If lacking risk assessments means, among other things, openmindedness about the possibilities for very long-term survival of the human race, then acceptance of the HR assumption radically curtails such liberality. Furthermore, in the form outlined in my introduction above, the argument employs no prior risk assessments. To settle this matter, I describe a procedure for deriving precise doom probabilities for cases with absolutely no prior risk estimates (see fn. 8). It is not true that the doomsday argument never pulls projections "out of the air."

There are other indications that Leslie has not fully faced up to the perversities of doomsday reasoning when applied to human destiny rather than to lotteries. For explicitness, begin with a lottery that has either 400 or 600 tickets. A valid random drawing yields ticket number 399. What should be concluded? A doomsdayer might reason as follows: "If the lottery has 400 tickets, the drawing is so unusual as to produce a ticket in the top 1/2% of all tickets, whereas if the lottery has 600 tickets, number 399 is much more reasonably placed in the 66th percentile. So randomly drawing ticket 399 strongly favors the larger pool." This reasoning is demonstrably incorrect. There is one chance in four hundred of randomly drawing ticket 399 from a pool of 400, and only one chance in six hundred of doing so from a pool of 600. By Bayes's theorem, you should adjust your estimates in favor of the smaller pool. This example not only exposes a peril inherent in this kind of reasoning, but highlights its most abiding peculiarity: when correct allowance is made for the randomness assumption, doomsday reasoning has an unalterable bias toward *earlier* doom.[8] The revision brought on by the HR assumption is not a redress of some optimistic bias in earlier estimates; it is an invariant feature of the argument that any and all revisions

---

[7] *The End of the World*, p. 204 (his italics); see also p. 213.

---

[8] A general formula for doomsday arguments: for greatest generality, I break the doomsday A_n scenarios into the finest possible alternatives by interpreting A_n to mean "doom will occur when the total cumulative human populations reaches n." (We then know $P(A_k)$ is zero for all k corresponding to people already born, since we know doom has not befallen us yet.) These A_ns can then be grouped and approximated, but it is an essential feature that "time left until doom" be reckoned through additions to the population; the HR assumption imputes randomness to individuals, not to time periods.

We let R stand for the fact that one's birth rank is r. By Bayes's theorem:

$$P(A_n|R) = \frac{P(R|A_n)P(A_n)}{\sum_{j=0}^{j=\infty} P(R|A_j)}$$

The crucial probability in this expression is $P(R|A_n)$, the probability of having birth rank r *given* that doom will occur at cumulative population n. If, quite reasonably, one supposes that the probability of having birth rank r remains the same as one that runs through various *later* doom scenarios, the argument stops in its tracks; for then $P(A_n|R)$ is just $P(A_n)$ and the prior probabilities are left unrevised. To secure doomsdayer conclusions, we need the HR assumption from which it follows that $P(R|A_n) = 1/n$ and that for any n:

$$P(A_n|R) = \frac{\frac{1}{n}P(A_n)}{\sum_{j=0}^{j=\infty} \frac{1}{j}P(A_j)}$$

Prior probabilities $\{P(A_n)\}_n$ shift to $\{(1/n)P(A_n)/\Sigma_j\ (1/j)P(A_j)\}_n$ under the impact of the HR assumption.

Suppose A_x and A_y are two doom scenarios with nonzero prior probability and such that upon application of the transformation A_x gains in likelihood and A_y loses, then:

shorten time left until doom. It matters not whether prior assessments predict early or late doom, whether current populations are large or small, growing or shrinking, or whether many or few have been born to date; there exists no constellation of circumstances under which a user of the argument concludes that prior assessments of doom ought to recede. If you are convinced that doom is a week away, the argument calls for earlier arrival; if you think that doom is likely in ten minutes, the argument presses for sooner. There is no room in this complex for the idea that original projections might err through earliness as well as lateness and that corrective action might require upward as well as downward revision. Can a procedure be valid that holds all other estimates of survival time to be too generous, irrespective of what they are or on what basis they were obtained?

Without the HR assumption, the proposed application of Bayes's theorem is trivial and fruitless, with the assumption the reasoning runs smoothly to its alarming conclusion. For a true lottery with an unknown quantity of tickets consecutively numbered starting with 1,

---

$$\frac{\frac{1}{x}P(A_x)}{N} - P(A_x) > 0 > \frac{\frac{1}{y}P(A_y)}{N} - P(A_y)$$

where $N = \Sigma_j \, (1/j)P(A_j)$. Since $P(A_x)$, $P(A_y)$ are positive, $1/xN - 1 > 0 > 1/yN - 1$; then $1/xN > 1/yN$ and $x < y$. Therefore, it is always earlier scenarios that gain likelihood from later scenarios.

To undermine the claim that the doomsday argument can only revise pre-existing risk assessments, I employ a 'prior' that renounces all such preconceptions in a thorough manner. (At this point, but only at this point, do I invoke techniques that are Bayesian in the controversial sense.) *A noninformative prior* can express *complete ignorance* as to, for example, the size of something (see James O. Berger, *Statistical Decision Theory and Bayesian Analysis* (New York: Springer, 1985), pp. 82-90). As applied to the magnitude of the final cumulative human population, such complete ignorance encapsulates not only freedom from risk assessments, but a thoroughgoing lack of hope, fear, prejudice, or opinion of any kind as to the size of this population (except for the belief that it will be finite).

Let r be the rank of an infant born now. For $n>r$ assign weight $w(A_n) = 1/n$ to A_n. (These weights are not probabilities since they necessarily sum to infinity.) Applying the above transformation to these weights, we obtain:

$$P(A_n) = \frac{\frac{1}{n^2}}{\sum_{j=r+1}^{j=\infty} \frac{1}{j^2}}$$

Unlike the original weights, these are true probabilities; note that the probability of doom scenario A_n depends exclusively on r and n.

the random drawing of a numbered ticket does indeed give us information about the probable size of the ticket pool. The issue in the doomsday argument, however, is whether such random lotteries constitute appropriate models—whether the HR assumption holds. In light of this, Leslie's copious urn and lottery examples can all be seen to be question begging; each one assumes equiprobability of sampling, precisely what must be established to validate doomsday reasoning. (In response to my objection to the assumption of random human sampling in the doomsday argument, Leslie produced additional urn and lottery analogies presupposing random sampling.*)

If one does not assume equiprobable selection, then natural examples cut the other way: suppose on each trial the *consecutive token dispenser* expels either 50 (early doom) or 100 (late doom) consecutively numbered tokens at the rate of one per minute. The sampler's task is to select a randomly ranked token *as it is expelled*. A moment's reflection reveals that this cannot be done without prior knowledge of the pool size for each trial; the sampler must contrive to choose randomly from a domain of 1 through 50 when the machine is going to expel only 50, and randomly from 1 through 100 in other cases, without knowing which cases are which. It makes not the least difference whether the pool size for each trial is decided by a deterministic or indeterministic process; the obstacle is ignorance, not indeterminateness. Without knowledge of the trial's termination, it is impossible to select random ranks.

Bayes's theorem cannot give us something from nothing. If the theorem is to tell us what consequence our birth rank has for doomsday, it can only be because we have fed in information as to what consequence doomsday has for our birth rank. Lacking knowledge of the latter, we cannot hope to apply Bayes's theorem profitably.

III. THE SHOOTING ROOM

Successive groups of individuals are brought into a room and given the same highly favorable wager, say, betting $100.00 that the "house," with fair dice, rolls anything but double sixes. (In the original formulation, losing players are shot, but this added gruesomeness, if nothing else, complicates the question of how one should bet.) Whenever the room occupants win their bets, ten times as many people are recruited for the next round. Once the house wins, the game series is over. So the house can truthfully announce before any games are played that, in spite of the highly favorable odds, at least 90% of all players will lose. The puzzle is that these bets appear to be both favorable and unfavorable: favorable because double sixes are rare, unfavorable because the overwhelming majority of players lose.

---

* "Doom and Probabilities."

Leslie analyzes the shooting room in detail.[18] His treatment closely parallels his approach to the doomsday argument. What are suspicious maneuvers in the case of the doomsday argument translate into demonstrable errors in the more explicit shooting-room story. Leslie considers two shooting-room variants, a deterministic and an indeterministic one, which we call the *D-series* and the *I-series*, respectively. In the D-series, for each round the one chance in thirty-six of losing is simulated by consulting the decimal expansion of π. The outcome of these games are completely predetermined, although players do not know these outcomes. In the I-series, identical odds are delivered by a device that is presumed to be truly indeterministic. Leslie concludes that in the I-series, players ought always to bet, since they have about a 97% chance of winning; but in the D-series players ought *not* to bet, since it is predetermined that at least 90% shall lose.

Leslie states that "there is nothing too paradoxical" in his approach (*ibid.*, p. 256). In fact, it requires only minor changes to the narrative to nudge this novel reasoning about the role of determinism into self-refutation. Here are two examples.

(1) Add a spectator who can bet in any of the D-series games. It is easy to show that the spectator should opt to play all games in the series, either for the customary reason that in each game there is a better than 97% chance of winning, or for the peculiar outcome-based reason that the spectator is thereby fated to a string of wins capped with a single loss. The most the spectator can lose is $100.00, but her expected winnings are measured in the thousands. By any accounting, the spectator ought to judge these games as highly favorable, and it is surely no argument against this that the precise moment at which her winning streak will end is preordained by the fortunes of π's expansion. Then for any game in the D-series, Leslie's reasoning dictates that the spectator's bet is favorable and the room player's bet is unfavorable, even though they both make the same bet on the same outcome of the same game.

Should it be maintained, for reasons I cannot fathom, that the spectators' statistics cannot be legitimately segregated from those of the room players, I also offer an example that requires no additional personnel.

(2) Play the I-series games first and record the results. These results, which are now a completely predetermined sequence just like

---

[18] *The End of the World*, pp. 235-36, 251-56.

π's decimal expansion, are then used for the D-series games. In other words, I-players and D-players bet on exactly the same "roll" in corresponding rounds, except D-players bet after the fact, but without knowledge of the earlier results. Once again, this reasoning collides with itself. The I-bets are favorable and should be taken, the D-bets are unfavorable and should not be taken, even though in this case these bets are on exactly the same outcomes of the same games.

My stance is that the shooting room is not a paradox at all; rather, it is a cogent line of reasoning alongside an utterly spurious one, masquerading as horns of a dilemma. The bets in all shooting-room games are highly favorable; no other conclusion is possible. The first thing to note is the solid dependence among outcomes for the room occupants in a single round—they all win or they all lose together. If five players lose independently at a *given kind* of wager, that may constitute, in a loose sense, five reasons not to play; but if a million players all lose the *same* wager at once, that constitutes, in the above sense, *one* reason not to play. (It has been remarked that an insurance company would go broke insuring all the players in the shooting room as though they had a 97% chance of winning. So might any insurance company that treats highly dependent events as though independent; for example, a diversified company that finds it rarely receives simultaneous flood claims and decides to insure everyone living on the banks of the Mississippi.) It has long been known that by successively increasing bet size in a sequence of unfavorable bets, one can theoretically obtain winning results;[11] this is the basis of various infamous doubling systems in roulette and other games. In the shooting-room game, it is the house that carries out the "multiplicative" betting scheme. Irrespective of whether it is a D-game or an I-game, each player ought to reason thus: 90% of all players will lose, but I have less than a 3% chance of belonging to that losing majority. This is no paradox; each player is prospectively likely to be in the minority, since she is prospectively likely to win and winning itself causes there to be enough subsequent players to guarantee the winner minority status.

The multiplicative betting system that the house exploits to secure a preponderantly losing final pool is not such as should make a difference to a player's decision. A player knows that the game is structured so that the majority ends up losing. This is to be achieved in one of two ways: (a) about a 3% chance the player will lose, along with enough others to assure the right preponderance of losers; (b) about a 97% chance the player will win, and enough other players will lose in some future round to secure the right proportion of losers. Surely, there is no reason to be found in (b) not to play. Nor should (a), a 3% chance that the player will lose along with many others, be an inhibiting factor. The bet is clearly favorable. (It may come as a relief to theoretical physicists that Leslie's alleged "destruction" of the many-worlds interpretation of quantum theory[12] relies on the same kind of reasoning that prompts him to judge betting even money against double sixes to be unfavorable.)

---

[11] For example, Richard A. Epstein, *The Theory of Gambling and Statistical Logic* (New York: Academic, 1977), pp. 52-56.

To explain how the bets can be so favorable for the players yet the house can be ultimately guaranteed an enormous profit, it may help to consider the consequences of playing in more than one shooting-room series. Call a sequence of ordinary shooting-room series a *megaseries*. Players for these consecutive game series are drawn from an unlimited store of numbered players. A player can only play once in any series; some series employ more players than others, but each new series begins with player 1 again, and so on. Let us examine the likely career of a particular player, *player n*. Lower numbered players in general play more often than higher numbered players, but it can be demonstrated that, if the megaseries continues long enough, player *n*, no matter how high her rank, eventually plays any preassigned number of games. Suppose she has played 100 times. How many times should we expect that player *n* had to play against double sixes? Recall she is any player and these are 100 distinct rolls of the dice or simulator. To suppose player *n* has faced double sixes about 90% of the time is to suppose the absurdity that 90% of all rolls of the dice in the games have been double sixes. One can only conclude that player *n* has had to contend with double sixes on average about three times out of the hundred, and that the more often she bets, the greater her expected net gain becomes. This confirms that the shooting-room games are highly favorable to players. It remains only to square the rising expected gain of each player with the fact that the house collects in a net sense enormous amounts from the store of players by the end of every game series. Examining the composition of net winners and net losers at virtually any stage of the megaseries, we would find the overwhelming majority of players who had bet more than once to be net winners. The house's advantage would be financed almost exclusively by vast numbers who had bet *once* and lost. As the megaseries continued, these net losers would have additional opportunities to play and, with freakish exceptions, would become net winners, while the ever expanding group of mostly one-time losers would drift ever higher in average rank. Note that the house's advantage depends crucially on the unlimitedness of the player store. With a finite store of players, no matter how large, the house would repeatedly find itself having to pay off all the players without there being any larger group from whom to reclaim the losses. (It is a feature of all multiplicative betting systems that they eventually fail disastrously.) Although theoretically assured of astronomical gains, any real house that undertook to run a shooting room would go broke absorbing bad bets while waiting for a huge payoff. It is most assuredly the house and not the player that makes foolish bets in this game.

---

[12] *The End of the World*, pp. 264-66.

IV ROOM AND DOOM

The doomsday argument relies on a subtle appeal to temporally reversed causation. I shall begin by exhibiting similar reliance in Leslie's treatment of the more explicit (deterministic) shooting-room story. Consider a variant with a limited number of rounds, say, five, even if no player loss occurs. We do not know the final frequencies for this variant with as much precision as in the standard shooting room, but we do know that roughly seven times out of eight the final pool consists exclusively of winners. These bets are without doubt highly favorable for the players. In Leslie's approach, although the version limited to five rounds is favorable for the player, the first five rounds of the standard version are unfavorable (all the games in that version are allegedly unfavorable). The sole difference between the two variants concerns what happens after the fifth round and to different players, that is, it is the possibility of later rounds with other players that allegedly renders the first five rounds of the standard series unfavorable.

The shooting-room story is structured so that the causal factors that determine whether a player wins or loses all factor through the roll of the dice; the composition of the final pool cannot make itself felt without reverse causation. Suppose you are a candidate player scheduled for a particular round. The dice frequencies influence your membership in the final player pool in at least two ways: (a) you are caused to become a player and hence to be a member of the final player pool by a player win on the *previous* round; (b) you are caused to be a majority or minority member of the final pool by the roll on your scheduled round; specifically, you will be a majority member only if the roll is double sixes. By contrast, the final pool frequencies are causally terminal; they influence no other event in the story. Your player status is caused not by your membership in the final pool, but by previous rolls of the dice. Hence, it is the dice frequencies and not the final loser frequencies that provide useful guides for action. (This can change, of course, if we expand or modify the context. Sampling randomly from the final pool yields a player with a 90% chance of having lost to double sixes. The final frequencies are in this case a reliable guide because the composition of the final population *influences* the sampling.)

Two peculiarities of the shooting-room story may be what delude some into supposing it reasonable to use causally terminal frequencies to govern earlier behavior. (1) The final frequencies are known beforehand with unusual precision; (2) the round on which you are caused to belong to the final pool comes before the roll of the dice on which you bet. These combine to make it appear that the determinate statistics of the final pool are already operative when the dice are rolled. A crucial point, and one that bristles with significance for the doomsday argument, is that it is an error to consider yourself a random or typical player until you lose your bet. Before then, you have only about a 3% chance of belonging to the 90% majority. (What lends an air of paradox is that an *unlikely* event makes a group of players *typical*.) This means *you should not consider yourself random until the game series is over.* We can highlight the analogy with doomsday reasoning by the following consideration: if a player about to bet were truly random among all players, then he would have better than a 90% chance of losing, which would mean that with high probability the game series would be over sooner than expected.

We return to the doomsday argument. The HR assumption has its effect because it stipulates a quantitative relationship between the probability of having your birth rank and the number of people who come after you. For the argument to be valid, this crucial probability cannot be based on your causal antecedents, but has instead to be based not only on how many were born before you but also how many are to be born after you. How it is possible in the selection of a random rank to give the appropriate weight to unborn members of the population? In presuming that unborn populations have somehow been factored into the current selection procedure, the doomsday reasoners tacitly presuppose retroactive causal effects. Suppose some crucial event prevents a holocaust in the year A.D. 2150 and doom is thereby delayed a thousand years. By doomsday reasoning, the probability of having a current birth rank is therefore lower than it would have been were this crucial event not to happen in A.D. 2150. This is true reverse causality, not some correlation artifact.

One can perhaps see this most easily in terms of a counterfactual formulation of causality; a counterfactual alteration to doomsday, for any reason whatever, is allegedly reflected in *current* probabilities. It is probably this tacit but largely unacknowledged reverse causation which is at the bottom of all the thrashing about over determinism. Unless the determinate numerosity of future generations somehow exerts the right influence on one's birth-rank selection, the argument is endangered; indeterminism threatens to render fuzzy and indistinct the needed retroactive future causes. Such retroactive influence is not inconceivable; however, its use further compromises the already tattered plausibility of the doomsday argument.

To summarize the major parallels between doomsday reasoning and the fallacious outcome-based approach to the shooting room: both are claimed to depend on determinism, yet in either case what is truly needed is not a determined future, but a retroactively influential one. In both, the assumption of the user's randomness is crucial and, in both, it has the consequence that the "game" is likely to be over sooner than otherwise expected. In the shooting room, this kind of thinking can be made to yield absurd and contradictory results; in the case of the doomsday argument, the elusiveness and ambiguity of human destiny help to conceal the argument's invalidity.

V. RANDOM IDENTITY

Examples such as the *consecutive token machine* illustrate the impossibility of choosing a random rank without knowledge of how many ranks exist overall (or without being able to interact with the population as in drawing from an urn). I believe the doomsdayer response to this would be that we may be unable to *select* random humans, but this does not prevent us from *being* random humans.

Doomsdayers precede rather freely from the item's membership in a class to its randomness in that class. The mere fact of being a likely candidate for random member in one class, however, often precludes a similar status in some other class. A random human is surely an exceptional vertebrate, and so on. Leslie proposes that a single human be taken as a random sample from the class of all humans *and* from the class of all mammals.[15] A random drawing from the class of all mammals, however, has only a minuscule probability of producing a human being; in fact, to have drawn a human from this class is excellent evidence that the drawing was *not* random. The instructions to select something that is both a random human and a random mammal are incoherent; if the item is to be a random mammal, we cannot also stipulate that it be human.

---

[15] "Doom and Probabilities," p. 491.

Leslie's expanded prescription that a selected item count as random in diverse classes to which it belongs leads to incorrect results; this can be readily seen by enlisting the former in an attempt to beat the consecutive token machine.[14] Suppose the sampler always selects a token from among the first 50 according to the principle that such a token is random in both smaller and larger pools. These fail as random choices for cases of late doom since, for that subset, all selected tokens are numbered 50 or lower, whereas randomly this should be true of only about half of them.

Whether or not we are likely candidates for the status of random human, we cannot be counted random mammals, random vertebrates, or random solid objects; we are too atypical as members of the last three classes.

Even though randomness in a class cannot be considered an automatic consequence of membership, there remains the possibility that randomness might arise from some ineffable element of human consciousness or human identity. Perhaps the best way to undermine this is to examine how the allegedly random user of the doomsday argument is elected. In order to derive significance from the value of the birth rank selected, we need to know something of what statisticians would call the *sampling density* of the selection procedure; the only conceivable way of accomplishing this is to examine how the sample is drawn. This is so obvious as scarcely to need mentioning, but the doomsdayers seem reluctant to take this step. In spite of their silence on what triggers selection, it seems clear that such sampling is undertaken only after the doomsday argument, or its central concept that we are randomly selected humans is discovered; in other words, the sampling density is that of waiting for a discovery. (Just as the randomness of selection from among all mammals is compromised by the precondition that the mammal be human, so the randomness of selection from among all humans who ever live is compromised by the precondition that the human be in possession of doomsday reasoning.) If the mean time until discovery of the doomsday argument is significantly shorter than the mean time until extinction of the race, then we would expect doomsday argument self-samples to be *early* members of the complete human population. In this case, the relative lowness of our birth rank loses its implica-

---

[14] *The End of the World*, pp. 256-63.

tion of early doom. To lend support to doomsday reasoning, we need a contrary assumption that the average time it takes to discover the argument is long compared to the average time it takes for the race to become extinct. The fact that the doomsday argument was independently discovered at least three times in recent years certainly detracts from the plausibility of this last assumption, but without it the strongest conclusion we can hope to draw is that we are likely to be of average rank among potential first discovers of the doomsday argument, which themselves may be biased toward earlier segments of humanity's complete history. In this case, discovery of the argument and the resultant sample rank imply nothing about the likelihood of early doom. In fact, the only conclusion about the timing of doomsday we can validly draw from our birth rank is the trivial one that doomsday does not occur until after the selection.[15]

We may have a variety of reasons to fear that the human race is doomed, but lowness of our own birth ranks cannot be counted among them.

WILLIAM ECKHARDT

Chicago, Illinois

---

[15] See my "Probability Theory and the Doomsday Argument," *Mind*, CII, 407 (1993): 483-88, here p. 487.

COMMENTS AND CRITICISM

DION'S FOOT

Suppose a certain man, Dion, has his foot amputated, and lives to tell the tale. That tale involves a well-known metaphysical puzzle, for most of us assume that there was, before the operation, an object made up of all of Dion's parts except those which overlapped with his foot—"all of Dion except for his foot," we might say, or Dion's *foot complement*. Call that object *Theon*. (Anyone who doubts that there is such a thing as Dion's undetached foot complement may imagine that 'Theon' is a name for Dion's undetached head. Surely, there is such a thing as Dion's head? Surely, Dion could, in principle, survive if his head were detached from the rest of him and kept alive?) It seems obvious that Theon, like Dion, continues to exist after the operation, for you cannot destroy an object merely by changing its surroundings—merely by removing something that was never a part of it. The puzzle, then (which might be called the *problem of undetached parts*), is how Dion and Theon are related after the operation.

The most common answer to this question is that Dion and Theon come to occupy just the same region of space and to be made of just the same matter after the operation. The next most popular answer is that Dion and Theon are made up of temporal parts, and while those of their temporal parts which "occur" before the operation only partly overlap, Dion and Theon have the very same postoperative temporal parts. Much as two roads can merge and have spatial parts in common, Dion and Theon merge and have temporal parts in common. Less popular accounts of the relation between Dion and Theon include relativizing identity to concepts or times, and denying that there is such a thing as Theon.

Michael Burke[1] has recently proposed an intriguing new solution (or resurrected an ancient one) to the problem of undetached parts. He argues that, despite appearances, Theon—Dion's foot complement—ceases to exist when Dion's foot is removed. The reason (leaving out many details) is this. Dion is a person both before and after the operation; Theon, however, is not a person. But every person is essentially a person, and every nonperson is essentially not a person. Thus, if Theon were to survive the operation, he would be a

---

[1] "Dion and Theon: An Essentialist Solution to an Ancient Puzzle," this JOURNAL, XCI, 3 (March 1994): 129-39. See also "Tibbles the Cat: A Modern Sophisma," *Philosophical Studies*, LXXXIV (1996): 63-74, and the article cited in footnote 2.

0022-362X/97/9405/260–65

nonperson at one time and a person later on, which is impossible. Hence, Theon must perish.

I want to explore two difficulties with this account. The first centers on the claim that Theon is not a person. Burke denies that Theon is a person on the grounds that personhood is a *maximal* concept, meaning that no proper part of a person is itself a person: since Theon is a proper part of Dion, and Dion is a person, Theon is not a person. But it is one thing to *say* that personhood is maximal, and quite another to explain how this could be so. How could being a part of another person prevent something from being a person itself?

Burke says that a person is an organism that, at some time during its existence, either has or has the potential to acquire certain mental capacities—rationality and consciousness, perhaps. Thus, if Theon is not a person, the reason must be either that he is not an organism, or that he is never rational and conscious (and could never become rational and conscious).

It is true that Theon is not an organism before the amputation. The reason why he is not an organism has to do with the fact that he is a part of another organism, Dion. But it would be surprising if Theon failed to be a person solely for that reason. If this were all that disqualified Theon from personhood, he would be a nonperson despite being psychologically just like a person. For Theon has the same brain as Dion, and thus presumably the same thoughts. It seems likely that Theon believes himself to be a person, since Dion believes this. The difference would be merely that Dion's belief is true and Theon's is false. In that case, most of the conscious, rational beings who think they are persons are mistaken—for if there is such a being as Theon, there are far more rational, conscious proper parts of human organisms than there are "whole" human organisms. If it is so easy to be mistaken about whether one is a person, I doubt that you and I have any reason to believe that *we* are persons. For all I know, I could be a proper part of a whole organism and therefore a nonperson. But if there are many nonpersons psychologically indistinguishable from persons, and if one could never know whether one was a person anyway, why should anyone *care* whether one was a person? This absurd view would deprive personhood of any philosophical interest.

So it is hard to see how Theon could fail to be a person, in any interesting sense of that term, simply because he is not an organism. The alternative is that Theon is not a person because he is not rational or conscious, and lacks even the capacity to become rational or conscious. This would be even more surprising than the view that Theon is rational and conscious but not a person, for Theon has, as a part, a fully developed human brain in perfect working order. The states of that brain mediate between the stimulations of Theon's sense organs, his movements, and other internal states of him in just the way that the states of Dion's brain do—Dion and Theon, in fact, share the very same brain states, sensory stimulations, and movements.

Burke chooses the second option, denying that Theon thinks or feels or acts. That is because thought, sensation, and action, like personhood, are maximal concepts: you cannot think or feel or act, Burke says, if you are a proper part of something else that thinks or feels or acts. This is an astonishing claim. How could the merely relational property of being a proper part of Dion prevent Theon from being rational or conscious? How could it prevent him from speaking English, or wanting a bath, or feeling pain? This can hardly be squared with current thinking in the philosophy of mind. A being that was exactly like Theon but *not* a part of another rational, conscious being—something that differed from Theon only in its surroundings—*would* be rational and conscious. Even if the *content* of one's intentional states can depend on one's surroundings, no current materialist theory of consciousness or intentionality is consistent with there being two physically indistinguishable beings, one of which was as rational and conscious as you or I, and the other of which was incapable of any thought or awareness at all.

Quite aside from this awkward problem, it is not clear how the *maximality principle* supports Burke's view. Burke uses the principle that no person (or thinking being) could be a proper part of another to argue that Theon is not a person because Dion is. But one could just as easily use that principle to argue that Dion is not a person because Theon is. As we have seen, Theon's intrinsic features, and the state of his nervous system in particular, would seem to make the premise that Theon is rational and conscious just as plausible as the premise that Dion is rational and conscious.

Even if you think that Dion's being a whole human being somehow makes him intrinsically better suited to being a subject of thought and consciousness than Theon, Dion may have other rivals that do not share Theon's defects. There may be a being made up of all of Dion's atoms and one more—an atom in the outer layers of Dion's skin that is nevertheless not a part of Dion, for example. Dion would be a proper part of that being, which we may call *Leon*. Thus, Dion would be a person only if Leon were not a person—only if Leon were not rational or conscious. But Leon would be just as "whole" as Dion is, and no intrinsic difference between them could make it the case that Dion was rational and conscious and Leon was not. If there were such a being as Leon, Leon himself would presumably be a part of a slightly larger being who would be equally well suited to be rational and conscious. Of course, if we continue to consider larger and larger beings in this way, we shall eventually get into beings that no longer seem to be rational or conscious: eventually we may come to consider a being composed of a human being and his socks, for example (though exactly what feature of such a being would prevent it from being rational or conscious will not be easy to say). Nevertheless, no *one* being that has Theon as a proper part will be better suited, intrinsically, to be rational and conscious than any other such being. Thus, if no proper part of a person is itself a person, it would seem that there are no persons at all. Likewise, if rationality and consciousness are maximal properties, then nothing is rational or conscious.

Someone might try to avoid this consequence of the maximality principle by denying that there is such a being as Leon, or anything else enough like Dion to count as a "whole human being." For each object that fails to be a person because it is a proper part of something else, there is never more than one candidate for the office of being the person of whom that being is a part. Dion simply has vague boundaries: there are atoms that are neither definitely parts of him nor definitely not parts of him. That would mean that there are atoms that do not compose anything (or at least atoms that do not *definitely* compose anything). Burke[2] has expressed doubts about whether just any atoms compose something. But if there are atoms that do not compose anything, there is little reason to believe that there is such an object as Theon in the first place. In fact, such an object would be quite problematic: if Theon existed, but there were nothing composed of just Theon's particles plus a few more, then nothing, it seems, could explain why Theon was made up of just *those* particles rather than those plus a few more—why he leaves off at the arbitrary place where he does.

Of course, any philosopher who believes that there are such things as foot complements, heads, and other undetached parts of human beings that have "thinking parts" faces the problem of what to say about those objects. If they are rational and conscious, as they appear to be, then none of us is ever really alone: we each share our thoughts, our meals, and our bed with a large number of other thinking beings, some of whom are parts of us and some of whom, perhaps, have us as parts. It is the contrast with this uncomfortable view that gives the maximality principle its attraction—even if we cannot understand how the principle could be true.

---

[2] "Preserving the Principle of One Object to a Place: A Novel Account of the Relations among Objects, Sorts, Sortals, and Persistence Conditions," *Philosophy and Phenomenological Research*, LIV (1994): 591-624, esp. p. 617. (The usual definition of 'compose' is that some things, call them the *x*s, compose something *y* just in case each of the *x*s is a part of *y*, none of the *x*s shares a part with any other of the *x*s, and every part of *y* shares a part with one or more of the *x*s.)

But there is an alternative to the maximality principle that is perfectly intelligible. It is the view, familiar to those who deal in the ontology of temporal parts, that we count people, or thinking things, as "one" whenever one of them is a part of the other.[3] Thus, even if there are, strictly speaking, a great many persons or thinking things sitting in your chair just now, when engaged in the ordinary business of life we say truly that there is just one, for we count people not by identity, but by an overlap relation. To say that you are alone is to say that every person in your vicinity is either a part of you or has you as a part—or, better, shares your thinking parts. This semantic hypothesis accounts for our ordinary judgments about how many people there are without making any problematic claims about what it takes to be a thinker. Since Burke's view relies on the assumption that proper parts of thinkers are not thinkers, however, it is not available to him.

Let me turn now to what appears to be a second difficulty with Burke's account of Dion and Theon. Dion, Burke says, is *essentially* a person, and Theon is essentially a nonperson; that is why Theon perishes when Dion's foot is detached from him. Dion and Theon belong to different sorts or substantial kinds. Part of the idea that *person* is a substantial kind is that the concept of a person determines the criterion of identity for all and only persons. Thus, if Theon were to survive the operation and come to be a person, he would start out having the persistence conditions for nonpersons (or, more precisely, the persistence conditions for some sort, the members of which are not persons) and end up with the persistence conditions for persons, which are presumably different from and incompatible with those. But, obviously, a thing cannot exchange its persistence conditions for new and incompatible conditions part way through its career.

The trouble is that after the operation, when he lacks a foot, Dion is intrinsically just like Theon was before the operation. (To be sure, Dion may have some scar tissue that Theon lacks. But it is hard to see how that could make any difference in the present context.) Thus, we have two intrinsically indistinguishable objects belonging to different sorts, with different and incompatible persistence conditions. The only difference between Dion as he is after the operation and Theon as he is before it is a difference in their surroundings: Theon has a foot attached to him and Dion does not. But that merely relational difference could hardly explain why they are able to survive different sorts of alterations.

---

[3] See, for example, David Lewis, "Survival and Identity," in A. Rorty, ed., *The Identities of Persons* (Berkeley: California UP, 1976). I discuss this view at greater length in "Why I Have No Hands," forthcoming in *Theoria*.

As it happens, this is much like Burke's reason for rejecting what I called the most popular view about the case of Dion and Theon, namely, that they come to coincide spatially and materially when Dion loses his foot.[4] Burke objects to coincident objects on the grounds that there would be no intrinsic difference between them which could account for their difference in sort, or their different criteria of identity. If a copper statue coincided with a statue-shaped piece of copper that would outlive the statue if it were crushed, there would be nothing to explain *why* the piece of copper but not the statue could survive being crushed—or, to put it differently, why the piece of copper is a piece of copper and not a statue.

This argument seems to provide an excellent reason for denying that Dion after the operation and Theon before it could belong to different sorts with different persistence conditions. If the piece of copper and the copper statue cannot have different persistence conditions without differing intrinsically, how do Dion and Theon manage to do so? To put it the other way round: If you can have two intrinsically indistinguishable objects belonging to different sorts in different places, why not two intrinsically indistinguishable objects of different sorts in the *same* place?

The problem is quite general, and has nothing to do with the particular claim that persons are essentially persons. If there is such an object as Theon, then Burke's theory requires that it cease to exist when Dion loses his foot because there is *some* sort of thing which Dion essentially is and which Theon essentially is not. Thus, Dion and Theon must have different persistence conditions. (Otherwise, it would be hard to see why losing a foot should destroy Theon but not Dion.) But again, Dion as he is after losing his foot is intrinsically just like Theon was beforehand, leaving nothing to account for their difference in persistence conditions.

Thus, even if the first difficulty can be overcome, and Burke's account is internally coherent, his view appears to have no advantage over its main traditional rival.

ERIC T OLSON

University of Cambridge

---

[4] "Copper Statues and Pieces of Copper: A Challenge to the Standard Account," *Analysis*, LII (1992): 12-17. See also "Dion and Theon" (135), and my "Composition and Coincidence," forthcoming in *Pacific Philosophical Quarterly*.

BOOK REVIEWS

The British Moralists and the Internal 'Ought': 1640-1740. STEPHEN DARWALL. New York: Cambridge University Press, 1995. xii + 352 p. Cloth $59.95, paper $18.95.

Stephen Darwall has written a fine book that is sure to change the way scholars approach the related issues of morality, practical rationality, and motivation in the development of British moral philosophy from Thomas Hobbes to David Hume. He accords these major figures careful treatment (as he does John Locke and Bishop Butler), but he deals in equal detail with such lesser-known thinkers as Nathaniel Culverwell and Richard Cumberland. His chapter on Lord Shaftesbury makes more philosophically interesting sense of that diffuse writer than any other I have seen. A long discussion of Ralph Cudworth uses unpublished manuscripts in the British Museum to paint an intriguing picture of his views that differs strikingly from received accounts.

Darwall's focus in all these writers is on the issue of internalism. The version of this doctrine that most interests him is *agent* or *existence internalism*, which holds that, if an agent ought to do something, then she must have, at least dispositionally, some motivation to do it. But Darwall insists on a crucial distinction. Some existence internalists—his eighteenth-century examples are Samuel Clarke, John Balguy, and Richard Price—take the relevant dispositional motive to be one an agent would have if she *knew* what she ought to do. Their internalism does not attempt to illuminate what "normativity itself" consists in, Darwall thinks, and indeed leaves this matter, the nature of what Clarke calls the "eternal fitnesses" of actions, quite mysterious. In contrast to these *epistemic internalists*, as we might call them, another view, *constitutive internalism*, "hoping to understand the nature of normativity itself, claims that the existence of a motive, perhaps of a certain kind or under certain circumstances, is (at least part of) what it is for a normative proposition to be true" (11).[1] It is on the development of this view, which aims to make the ought itself "internal" to the agent, which Darwall concentrates.

He draws some further distinctions. Empirical naturalists (Hobbes, Cumberland, sometimes Locke) start by trying to fit the required motive into a (largely) naturalistic world picture, whereas autonomists (Cudworth, Locke in other passages, but paradigmatically Shaftesbury and Butler) start with the idea that the motive must arise from an agent's autonomous deliberation. These classifications are by Darwall's admission rough: he is tracking what he regards as sometimes only anticipations, and in any case never fully-worked-out versions of these positions. Even by these rough standards, however, not everyone gets classified as an internalist. Francis Hutcheson, who figures here mainly as a foil for his predecessors and contemporaries, does not. Hume is a trickier case. He does not seem a constitutive internalist, but Darwall provides a speculative but meticulously defended reading that would make him one, not in his discussion of moral obligation generally, but in his otherwise very puzzling discussion of justice. I found this discussion one of the most interesting in the book.

---

[1] Darwall uses the term 'constitutive internalism' only once (11n); except when he needs more distinctions he just calls it "internalism," as I shall also. He uses no special term for epistemic internalism, and sometimes even refers to Clarke, Balguy, and Price as externalists (for example, 19, 325).

Darwall does not entirely reject L. A. Selby-Bigge's familiar distinction between rationalist and sentimentalist traditions in response to Hobbes. But his primary aim is to establish that there is a far more important internalist tradition which cuts across this divide, and which culminates in an autonomism that in its more mature formulations anticipates Immanuel Kant (an offstage hero in this plot). He also hopes to contribute to contemporary debate about the foundations of ethics, using these figures to illustrate (especially for any of us who find Kant a bit opaque) what an internalist, and especially an autonomist, position might look like, and why it might seem attractive.

He seems to me largely successful in the first of these projects. Others will no doubt join me in resisting this or that claim of detail (perhaps different details). But this is a work of prodigious scholarship guided by keen philosophical intelligence, and Darwall's cumulative case, both about individual thinkers and about this century of moral thought, is impressive. About every writer whom he treats at any length, he has changed my mind on at least some significant points of interpretation, almost always in ways that support his case that there is more internalism here, playing a more central role, than most have supposed. It would be unwise, I think, to discuss any of these writers without considering Darwall's readings; and it will be hard to take seriously any subsequent attempt to deal with this important period as a whole which does not carefully come to terms with this book.

Darwall's second project is also of great interest, but seems to me to face more problems (and for reasons that at least cast a shadow over the first). My most important worry is that Darwall's crucial distinction between epistemic and constitutive internalism seems inadequately drawn. Constitutive internalists say that, if an agent ought to do something, then there are specifiable circumstances in which she would be motivated to do it. But Darwall's epistemic internalists do not reject this formula: they just take the relevant circumstances to be ones in which the agent knows what she ought to do. It is tempting to suppose that we are intended to supply the obvious qualification: constitutive internalists are those who say that, if one ought to do something, then there are specifiable circumstances, *other* than those in which one knows that one ought, in which one will be motivated to do it. But this amendment does not quite work. It covers Darwall's empirical naturalists and some of his autonomists well enough. For example, he finds in Locke's *Essay* the suggestion that what one is fundamentally obligated to do is whatever one would be motivated to do after full reflection on the facts; and apparently for Locke, as more recently for Richard Brandt, this reflection is supposed to be value free and, more importantly here, obligation free (19, 168). It is not thought about what one ought to do. The revised formula runs into more difficulty, however, when applied to Darwall's favorites among the autonomists, the ones who most clearly anticipate Kant. For these philosophers hold what Darwall calls a *normative theory of the will* (18-20), by which he means that according to them the motivation that determines what one ought to do is motivation that would result precisely from autonomous reflection on normative questions, questions about what one ought to do. He appears to read these philosophers as cognitivists, moreover, for whom such a conclusion, properly arrived at, amounts to knowledge. But this means that they agree, in their internalism, with the epistemic internalists that, if one ought to do something, then one will be motivated to do it if (in the right way) one knows one ought; it also means that the revised definition of constitutive internalism threatens to exclude them.

How, then, to distinguish these autonomists from the merely epistemic internalists? Darwall suggests several intriguing ideas, though none seems to me fully satisfactory. (1) One is that the epistemic theorists make moral thought passive (as emphasized in their speaking of moral *perception*), whereas for the autonomists it is active. But this contrast seems to me quite elusive. Both parties treat moral thought as motivating, so as active in that respect. There are passages in Price that indicate that moral issues may be difficult and so require a lot of active thought (whereas Butler by contrast can sound quite complacent on this score). Darwall finds in Cudworth the Cartesian-sounding view that all judgment, including practical judgment, is itself voluntary (135-39), but does not trace this doctrine into his paradigmatic autonomists, Shaftesbury and Butler.

(2) Another suggestion concerns the purported *location* of that normative fact, of which the knowledge motivates. Darwall frequently calls Clarke, Balguy, and Price "Platonists," and so presumably sees Clarke as locating his eternal fitnesses of actions in a Platonic heaven: not within the agent, at any rate, but somewhere "external." All his autonomists, by contrast, are like other constitutive internalists in urging that in deliberating about what we ought to do we must look inward. Darwall is right about something here. His favored autonomists are indeed distinguished by insisting that much basic normative thought has a second-order character: it is not directly about what actions we ought to perform, but instead about which of our motives or motivational faculties ought to take precedence. Thus, Butler has agents considering whether to follow self-love or particular passions; and, as Darwall shows, Shaftesbury and Cudworth emphasize similar questions. This is an important difference. But it is not clear to me how, by itself, it warrants saying that the autonomists make the ought itself internal, or that they differ from the epistemic theorists in attempting an account of what the ought itself consists in, rather than of what effect knowledge of it has. If one of Butler's agents sees that self-love should take precedence over particular appetites, the self-love and the appetites are indeed internal to the agent. But—especially since, on Darwall's reading, Butler and the other best autonomists resist any reductive naturalistic (for example, psychological) account of what makes it obtain—where is this normative fact about their relation located? Why not in that same Platonic heaven? Perhaps we should not take this talk of the "location" of normative facts seriously; but if we do, we need more reason than this for seeing autonomists as putting them "inside" the agent.

(3) That further reason is supposed to be supplied, I suspect, by one more consideration Darwall recurs to several times. Our problem arises from the fact that epistemic internalists and some autonomists agree on a certain formula: that, if an agent ought to do something, then proper reasoning about what she ought to do would lead her to think she ought to do it, and would motivate her to do it. But, Darwall suggests, they disagree about what is explaining what in this formula. The epistemic theorists see the normative facts as existing independently; what makes normative reasoning proper is that it discovers and is motivated by them. The favored autonomists, by contrast, start with an independent notion of proper normative reason-

ing, and take the normative facts to be determined, at least in part, by what proper reasoning would decide they were (20, 206). (As best I can judge, if they are fully autonomist, they also drop the 'at least in part'.) Darwall's autonomists are, in short, constructivists.

(4) This locates their position on a philosophical map, but does little yet to clarify it or to keep it from seeming (as it does to many) paradoxical. What might relieve this impression, especially in a book like this one, is some help by illustration: What would such a view look like as developed by a serious adherent? What Darwall provides from this direction, however, appears to me only to confirm the suspicion of paradox.

Of course, none of these writers is in his eyes a perfect autonomist, but (no doubt the least surprising of Darwall's historical judgments) Butler clearly comes closest. As Darwall knows, however, there is a problem in fitting together some of Butler's most autonomist-sounding views. The most prominent of these is that conscience is supreme, and so ought to be followed in any case of motivational conflict. Other of his views, however, including ones he seems to appeal to in arguing for the supremacy of conscience, appear to render that doctrine superfluous. If conscience is guided only by the standard that superior faculties should be followed (plus information about which are superior), then it will not matter at all to the deliverances of conscience whether it is a superior faculty itself.[2]

Darwall extricates Butler from this difficulty by arguing that it is wrong about the standards conscience employs. In addition to enjoining us to follow superior principles, he maintains, conscience also favors certain actions as fit "in themselves" in Clarke's sense. This is indeed a possible solution, with some textual support. An obvious question for Darwall, however, is how it is supposed to leave Butler essentially an autonomist, or to help display autonomism as an attractive position. For a fully autonomist Butler would surely have to eschew any reliance on perception of normative eternal fitnesses, the core of that puzzling epistemic internalism from which autonomists like him are supposed to be helping us escape.[3] But, on

---

[2] This is too quick: for a fuller explanation, see my "Nature and Conscience in Butler's Ethics," *Philosophical Review*, LXXXV (1976): 316-56. Darwall addresses this problem (249-61).

[3] On Darwall's reading of Butler, these fitnesses do not impose "obligations": only conscience creates those. But this cannot mean, as Darwall appears to suggest (254, 279), that the fitnesses are not normative in the sense of providing reasons. If they provided conscience no reason for choice, flouting a conscience responding to them would involve no disrespect for it as a superior, rational faculty.

Darwall's own account, a Butler without eternal fitnesses would again be a Butler for whom the supremacy of conscience is a superfluous doctrine. This difficulty seems to me to leave it quite unclear what a fully autonomist position would amount to.

Darwall's book is far richer in both historical and philosophical insight than a review can easily indicate. It is required reading for anyone interested in its period and will be a stimulating education for anyone thinking about its issues.

NICHOLAS L. STURGEON

Cornell University

NEW BOOKS: TRANSLATIONS

ARISTOTLE: *Introductory Readings*. TERENCE IRWIN and GAIL FINE, trans. Indianapolis: Hackett, 1996. xvii, 359 p. Cloth $34.95; paper $9.95; and *On the Heavens I and II*. STUART LEGGATT, trans. Warminster, UK: Aris and Phillips, 1996. vii, 273 p. Cloth $49.95, paper $24.95.

BLUMENBERG, HANS: *Shipwreck with Spectator: Paradigm of a Metaphor for Existence*. STEVEN RENDALL, trans. Cambridge: MIT, 1997. 124 p. Cloth $19.00.

BRENTANO, FRANZ: *Descriptive Psychology*. BENITO MUELLER, trans. New York: Routledge, 1996. xxvi, 198 p. Cloth $59.95.

DERRIDA, JACQUES: *Archive Fever: A Freudian Impression*. ERIC PRENOWITZ, trans. Chicago: University Press, 1997. vii, 111 p. Cloth $17.95.

DESCARTES, RENÉ: *Discourse on Method and Meditations on First Philosophy*. DAVID WEISSMAN, ed. New Haven: Yale, 1996. x, 383 p. Cloth $35.00, paper $15.00.

HAMACHER, WERNER: *Premises: Essays on Philosophy and Literature from Kant to Celan*. PETER FENVES, trans. Cambridge: Harvard, 1996. iv, 393 p. Cloth $55.00.

HYPPOLITE, JEAN: *Introduction to Hegel's* Philosophy of History. BOND HARRIS and JACQUELINE BOUCHARD SPURLOCK, trans. Gainesville: Florida UP, 1996. xix, 88 p. Cloth $29.95.

JOAS, HANS: *The Creativity of Action*. JEREMY GAINES and PAUL KEAST, trans. Chicago: University Press, 1996. x, 336 p. Cloth $49.00; paper $19.95.

KANT, IMMANUEL: *Critique of Pure Reason*. WERNER S. PLUHAR, trans. Indianapolis: Hackett, 1996. lix, 1030 p. Cloth $55.00, paper $21.95; and *Practical Philosophy*. MARY J. GREGOR, ed. and trans. New York: Cambridge, 1996. xxxiii, 668 p. Cloth $75.00; and *Religion and Rational Theology*. ALLEN W. WOOD and GEORGE DI GIOVANNI, eds. and trans. New York: Cambridge, 1996. xxiv, 518 p. Cloth $70.00.

MAY, REINHARD: *Heidegger's Hidden Sources: East Asian Influences on His Work*. GRAHAM PARKES, trans. New York: Routledge, 1996. xviii, 121 p. Cloth $49.95; paper $16.95.

PLATO: *Parmenides*. R.E. ALLEN, trans. New Haven: Yale, 1997. xiv, 351 p. Cloth $45.00.

POPPER, KARL: *In Search of a Better World: Lectures and Essays from Thirty Years*. LAURA J. BENNET, trans. New York: Routledge, 1996. x, 245 p. Paper $17.95.

RICOEUR, PAUL: *A Key to Edmund Husserl's* Ideas I. BOND HARRIS and JACQUELINE BOUCHARD SPURLOCK, trans. POL VANDEVELDE, ed. Milwaukee: Marquette, 1996. 176 p.

SCHLEIERMACHER, FRIEDRICH: *On Religion: Speeches to Its Cultured Despisers*. RICHARD CROUTER, ed. and trans. New York: Cambridge, 1996. xlv, 128 p. Cloth $54.95; paper $20.95.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 6, JUNE 1997

CALCUTTA UNIVERSITY LIBRARY

THE RIGHT AND THE GOOD*

There is a story about the right and the good which strikes me as very attractive.[1] Although there are gaps in it, I think its structure must be correct. But first some background.

The story about the right and the good which comes down to us from G. E. Moore—I shall call it *Moore's story*—is familiar enough, and many people find *it* very attractive. Many people nowadays reject it, but even among them, many understand perfectly why its friends resist invitations to reject it: that is because they themselves accept its first two parts and concede that it is hard to supply a clear rationale for rejecting its concluding third part.

Moore's story begins with the good. Some things are good, Moore said, and some things are not good; so there is such a property as goodness—all good things have it and all things that are not good lack it.[2] Moore had much to say about which things do and which things do not possess the property goodness, but for a reason that

---

* An early draft of this article was presented at the Chapel Hill Colloquium in 1995, and later drafts were presented as the Hagerstrom Lectures for 1996 at the University of Uppsala, and at a symposium at University College, London; I am grateful to the participants for their comments. The later drafts were written at the Centre for Advanced Study in Oslo; I am grateful to the Centre for providing me with support and very pleasant surroundings in which to work. Many people gave me helpful comments along the way: I thank Alexander Byrne, James Higginbotham, Ralph Wedgwood, and, in particular, Catherine Elgin, who commented on several drafts.

[1] I shall be trying to improve on the story I told in chapter 8 of Gilbert Harman and my *Moral Relativism and Moral Objectivity* (Cambridge: Blackwell, 1996).

[2] Moore said in *Principia Ethica*: "Ethics is undoubtedly concerned with the question what good conduct is; but, being concerned with this, it obviously does not start at the beginning, unless it is prepared to tell us what is good as well as what is conduct. For 'good conduct' is a complex notion: all conduct is not good; for some is certainly bad and some may be indifferent. And on the other hand, other things, besides conduct, may be good; and if they are so, then, 'good' denotes some property, that is common to them and conduct; and if we examine good con-

0022-362X/97/9406/273-98

will emerge, I shall not summarize his views on this matter. The second part of the story flows from the first: there being such a property as goodness, there is also such a relation as being better than, or *betterness.* (Not every property has a comparative, but this one surely must.) Moore's story then concludes: the right is analyzable in terms of the relation betterness. Thus, for it to be the case that Alfred ought to do a thing at a time is for it to be the case that the world will be better if he does the thing than it will be if he does any of the other things it is open to him to do at the time.

There are at least three things that make this story seem attractive. In the first place, it is wonderfully simple. It counts in favor of a theory if it supplies a simple picture of what had initially looked like chaos. Second, it does seem to us that there is such a thing as one thing's being good and another's not being good, and that there is such a thing as the world's being better if things go this way rather than that. We certainly say *words* like this often enough. Third, and most important, it is hard to see how it could be perfectly all right to fail to do what would make the world go better than it otherwise would. If you really would make the world go worse if you did a thing, then are you not called on morally to refrain from doing it? What else is there for the right to turn on than how good the world will be if you act in this way rather than that?

As I said, many people nowadays reject Moore's story. The ground most commonly given for doing so is that accepting the story yields intuitively implausible conclusions about the right. Suppose that Alfred is under the following threat by the Mafia: kill Bert, or we shall kill Charles, David, and Edward. Moore's opponents say: under those conditions, the world will surely be better if Alfred kills Bert than if he leaves the Mafia to kill the three others. After all, deaths are surely bad, and three of them surely three times as bad as one. But, they go on: moral intuition delivers, firmly, that it would be wrong for Alfred to kill Bert. In short, what is wrong with Moore's story lies in its concluding part. All is well with its two earlier parts, which tell us about goodness and betterness; the trouble lies only in the story's then going on to declare that the right is analyzable in terms of betterness.

---

duct alone of all good things, then we shall be in danger of mistaking for this property, some property which is not shared by those other things ..."—Thomas Baldwin, ed. (New York: Cambridge, 1993, rev. ed.), p. 54. Thus, good conduct is what has the two properties goodness and being conduct, and other things besides good conduct have the property goodness: goodness is the property that all good things have in common.

Moore's friends—and he still has a good many—reply: "But look, you have conceded that the world really will be better if Alfred kills Bert; so how could it possibly be wrong for him to do so? What is there that could be thought to make his doing so wrong, given that the world will be better if he does it?" This, in effect, asks "What's your alternative story?" and expresses skepticism about the possibility of any plausible alternative. There is much to be said for that skepticism, for it *is* hard to see what could make Alfred's killing Bert wrong if it is not bad, and then hard to see how the Mafia's killing the three other people could fail to be at least three times as bad.

This dispute still goes on in moral philosophy, and what is puzzling about it is that a deeper objection to Moore's story not only is available but has been available for many years now. Peter Geach (in 1956), Paul Ziff (in 1960), and G. H. von Wright (in 1963)[3] gave the excellent advice that we should *look and see* how the word 'good' is actually used. They showed, conclusively, that it does not function in the way in which adjectives like 'visible' and 'happy' do. In saying 'That's good', we are not ascribing a property goodness—indeed, there is no such thing. Curiously enough, this idea was not picked up by moral philosophers, and brought to bear on Moore's story, for some twenty years. My impression is that Philippa Foot's article "Utilitarianism and the Virtues" (1983)[4] is the first work by a moral philosopher to declare that since there is no such property as goodness, Moore's story is incoherent from the outset.

It should really have been clear back in 1903 that there was something fishy about Moore's story. At the beginning of *Principia Ethica*, Moore says that the question he will be addressing himself to in what follows is the question 'What is good?', and he rightly thinks that we are going to need a bit of help in seeing exactly what question he is expressing in those words. He proposes to help us by drawing attention to a possible answer to the question he is expressing—that is, to something that would be an answer to it, whether or not it is *the* or *a* correct answer to it. Here is what he offers us: "Books are good." Books are good? What would you mean if you said 'Books are good'? Moore, however, goes placidly on: "though [that would be] an answer obviously false; for some books are very bad indeed" (*op. cit.*, p. 55). Well, some books are bad to read or to look at, some are bad for use in teaching philosophy, some are bad for children. What sense could be made of a person who said, "No, no, I meant that some books are just plain *bad things*?" There is something weird about that passage in *Principia*, and it is puzzling that this was not noticed at the time.

---

[3] See Geach, "Good and Evil," *Analysis*, XVII (1956): 33-42, reprinted in *Theories of Ethics*, Philippa Foot, ed. (New York: Oxford, 1967); Ziff, *Semantic Analysis* (Ithaca: Cornell, 1960); von Wright, *The Varieties of Goodness* (London: Routledge, 1963).

[4] Foot's article was first published in *Proceedings and Addresses of the American Philosophical Association*, LVII (1983): 273-83. A revised version appeared in *Mind*, XCIV (1985): 196-209, and was reprinted in *Consequentialism and Its Critics*, Samuel Scheffler, ed. (New York, Oxford, 1988). I attribute the declaration to Foot's article and not to Foot herself, since while I continue to believe that the article can be interpreted as making it, Foot has recently said that she did not mean to do so, and she must be allowed final say as to her intentions.

It is worth mention that the idea at least *lurked* in von Wright. That exceedingly interesting book unfortunately had no impact on moral philosophy at the time; I attribute that fact to the seductiveness of Moore's story.

In any case, it is should be clear enough by now that all goodness is goodness in a way, and that, if we do not know in what way a man means that a thing is good when he says of it 'That's good', then we simply do not know what he is saying of it. Perhaps he means that it is good to eat, or that it is good for use in making cheesecake, or that it is good for Alfred. If he tells us, "No, no, I meant that it is just plain a *good thing*," then we can at best suppose he is a philosopher making a joke.

The same is true of betterness: it, too, is always betterness in a way. People do say the words 'This is better than that', but what they mean is always that the first thing is better to eat, or better for use in making cheesecake, or better for Alfred, and so on.

But if there is no such property as goodness, and no such relation as betterness, then, a fortiori, the right is not analyzable in terms of the relation betterness, and we need not appeal to moral intuition to justify rejecting Moore's story. In sum, what is wrong with Moore's story is not merely its concluding part but its very beginning.

We need a new story about the right and the good, then. Here is the one that strikes me as attractive.

I

It begins with what I shall call the *first-order ways of being good*. I mentioned being *good to* eat; there are also being good to look at, or listen to. I mentioned being *good for use in* making cheesecake; there are also being good for use in hammering in nails, or planting bulbs. I mentioned being *good for* Alfred; there are also being good for Alfred's lawn mower, or Alfred's apple tree. There are others, too: being *good at* hanging wallpaper, playing chess, or singing, being *good in Hamlet* (the play), being *good as* Hamlet (the Prince), being *good with* children, and so on. Let us follow a practice common among linguists in using the term 'adjuncts' to refer to the expressions 'to eat', 'for use in making cheesecake', and so on, which I added to 'good' to get 'good to eat', 'good for use in making cheesecake', and so on. In general, then, whenever we predicate an expression of the form 'good plus adjunct', we ascribe a first-order way of being good.

Perhaps the following will suggest itself: so after all there really is such a property as goodness!—it is the disjunction of the properties being good to eat, being good for use in making cheesecake, being good for Alfred, and so on. Or more compactly, that being good is being good in at least one of those ways. No doubt there is a property that is the disjunction of all of those properties. But it is an uninteresting property, for everything has it: everything is good in one or other of those ways. (If you find an example that strikes you as good in none of those ways, then it is sure to be good for use in a philosophical discussion of goodness.) It is therefore of no interest to any friend of Moore's story. Moreover, it is not what anybody ever means to be ascribing to a thing in saying 'That's good', so its title to be called "goodness" is at best dubious.

What about what we ascribe when we predicate an expression of the form 'good *K*'? Geach had said that 'good' is an attributive adjective, meaning by that: '*X* is a good *K*' is not equivalent to '*X* is good' and '*X* is a *K*'. Compare 'big', which is also attributive in that '*X* is a big *K*' is not equivalent to '*X* is big' and '*X* is a *K*'. (Geach said: by contrast, an adjective such as 'red' is predicative, for '*X* is a red *K*' is equivalent to '*X* is red' and '*X* is a *K*'.[5]) Indeed, he said something stronger, namely, that 'good' is "essentially an attributive adjective," meaning by that:

> Even when "good" or "bad" stands by itself as a predicate [as, for example, in "That is good"], and thus is grammatically predicative, some substantive has to be understood; there is no such thing as being just good or bad, there is only being a good or bad so-and-so (*op. cit.*, p. 65).

Thus, 'good' is like 'big' not merely in being an attributive adjective in the sense Geach indicated, but also in the following respect: just as there is no such thing as being just big, but rather only being a big so-and-so, so also there is no such thing as being just good, but rather only being a good so-and-so—in both cases, "some substantive has to be understood."

If this stronger claim were true, then we could (again) say that there is such a property as goodness: it is a relative property like big-ness. (As being big is being a big *K*, so also being good is being a good *K*.)

---

[5] 'Red' was not in fact well chosen for Geach's purposes, since 'red' is heavily context dependent: what we ascribe to an apple when we say 'It's red' is different from what we ascribe to the paint in a certain can when we say 'It's red'. Better choices would have been 'visible' or 'happy', which I mentioned above, or any of the following: 'sauteed' ('pureed', 'boiled'), or 'poisonous' ('nauseating'), or 'fermented' ('carbonated'), or 'acidic' ('alkaline'), and so on.

But the stronger claim is not true. The weaker claim is true: 'good' certainly is an attributive adjective in the sense Geach indicated, for '*X* is a good *K*' is not equivalent to '*X* is good' and '*X* is a *K*'. Moreover, it is certainly right to say there is no such thing as being just good. But it is not right to say that "some substantive has to be understood." If a man says "That's good," he *may* mean the likes of "That's a good book" or "That's a good banana"; but he may instead mean the likes of "That's good for use in making cheesecake" or "That's good for Alfred," and here no substantive needs to be understood. (In fact, for what *K* could it at all plausibly be thought that being good for use in making cheesecake is being a good *K*? For what *K* could it at all plausibly be thought that being good for Alfred is being a good *K*?)

I suggest, in addition, that our understanding of assertions of sentences of the form '*X* is a good *K*'—like our understanding of assertions of 'That's good'—rests on our knowing, from the context of utterance, in which of what I called the first-order ways of being good the speaker means that the thing is good.[6] I should think that most assertions of 'That's a good book' are made in contexts in which what is in question is which book to read, and the speaker is therefore likely to mean that the thing is a book that is good to read. But if what we had asked for was advice about which book to use in teaching elementary logic, the speaker is likely to mean, instead, that the thing is a book that is good for use in teaching elementary logic. In still other contexts, the speaker is likely to mean that the thing is a book that is good for children to read, or that it is a book that is good to look at. Similarly for 'That's a good banana': in most contexts, the speaker is likely to mean that the thing is a banana that would be good to eat, but where what we wanted was a banana for use in the display of fruit on the dining room table, he is likely to mean that it is a banana that is good to look at.

The context has to tell us what a speaker means by 'That's a good *K*' (as what a speaker means by 'That's good') or we simply do not know, for the expression 'good *K*' leaves this open. The point emerges most starkly when we look at odd expressions of the form 'good *K*'. Suppose someone calls a thing a good corpse.[7] What on earth does the speaker mean? I have no idea, because I have no idea what he means to be saying the corpse is good for use in, or good to, or good for, and so on. ('Good state of affairs' and 'good event' are unlike 'good corpse': it is not that we have no idea what a speaker might mean by predicating them, but rather that there are too many possibilities—the speaker might mean that the thing is a state of affairs or event that is good for him and his friends, or for people generally, or that it is good for use as an example in a class in philosophy of history, and so on and on.)

---

[6] Or in which of what I shall in the following section call *second-order ways of being good* the speaker means that the thing is good, for that, too, is a possibility.

[7] That lovely example comes from Ziff.

So just as there is no such a thing as being just good, there is no such thing as being just a good *K*—for example, just a good book or just a good banana. A fortiori, the instances of being a good *K* are not themselves among the ways of being good.

II

In some contexts, a person who says 'That was a good act' is likely to mean it was an act that was good to look at; but in other, perhaps more common contexts, a person who says this is likely to mean it was an act that was just, or generous, or kind, or considerate, and so on. An ascription of being just or being generous and so on is praise; and it seems clear that these are further ways in which a thing can be good. But they are intuitively not first-order ways of being good. Certainly, '*X* is just' is not equivalent to anything of the form '*X* is good plus adjunct', and similarly for the rest. The story I am telling says that these are *second-order ways of being good.*

There are other second-order ways of being good besides the moral. There are the aesthetic: being graceful, striking, dazzling, serene, witty, and so on. There are also being charming, elegant, sophisticated, and so on. (Are those aesthetic properties?) What about being healthy? I should think that it, too, is a way of being good. (I should think that being healthy is being in good health.) Is '*X* is healthy' equivalent to anything of the form '*X* is good plus adjunct'? Presumably not, and I therefore include it among the second-order ways of being good.

The context of an assertion of 'That's good'—as of a sentence of the form '*X* is a good *K*'—may make it likely that the speaker means that the thing is good in a first-order way ('good for use in making cheesecake', 'good for Alfred') *or* that it is good in a second-order way ('just', 'graceful').

Why *second-order*? My suspicion is that while none of the ways of being good that I here call second-order is *simply* reducible to a first-order way of being good, they rest on first-order ways of being good. How? I shall restrict myself to the moral second-order ways of being good, since it is only those which are relevant to the story I am telling.

Two preliminaries. First, it is not merely some acts that are just, generous, and so on: some people are so, too. But it is surely plain enough that we should not take the property of being just to be the "what is in common" to all just entities, people as well as acts, the possession of which marks them all as just. (Compare Moore on goodness in the passage I quoted in footnote 2 above.)

A helpful model is Aristotle's example of health. Some people are healthy, and so are some foodstuffs; but being healthy is not the property had in common by all healthy things—rather, being healthy is what all healthy *Xs* have in common, healthy *Ys* being healthy only derivatively, in the appropriate sense. Surely, the *Xs* here are people, and the *Ys* foodstuffs; that is, surely being healthy is what all healthy people have in common, healthy foodstuffs being healthy only derivatively, in the sense that eating them is conducive to a person's becoming or remaining healthy.

So similarly, we may take it that being just is what all just *Xs* have in common, just *Ys* being just only derivatively, in the appropriate sense. But which are the *Xs* *here*? Some philosophers say that acts are the *Xs*: thus they say that being just is what all just acts have in common, just people being just only derivatively, in the sense that they are prone to performing just acts. Other philosophers say that people are the *Xs*: thus they say that being just is what all just people have in common, just acts being just only derivatively, in the sense that they are acts that a just person would perform. (I take this to be Aristotle's view.) The notions "just act" and "just person" are presumably interdefinable: Which way should analysis proceed? I am myself in the former group; that is, I think that an act's being just is metaphysically prior, and a person's being just metaphysically secondary, and I shall assume that that is so. But I shall not argue for this idea, since the story I am telling does not rely on it. Given interdefinability, the story I am telling could as well have been told the other way round.

In sum, I shall take the noun phrase 'being just' to refer to what all just acts have in common, just people being just only derivatively, in the sense that they are prone to performing just acts.

But I shall follow a common usage according to which the noun 'justice' refers to a character trait possessed by people, namely, the character trait that consists in proneness to performing just acts. (Perhaps it need not be said, though it will be, that this proneness to performing just acts includes proneness to avoiding unjust acts, since it includes proneness to performing just avoidances.)

Similar questions arise in respect of being generous, being kind, and so on.[8] I shall take the noun phrase 'being generous' to refer to what all generous acts have in common, generous people being generous only derivatively, in the sense that they are prone to performing generous acts. I shall take the noun 'generosity' to refer to the character trait that consists in proneness to performing generous acts. Similarly for 'being kind' and 'kindness', and so on.

A second preliminary is required because the list of moral second-order properties that I gave trails off into 'and so on'. What fixes what does and does not belong on the list? I shall have more to say about this in the following section. For the moment, I mean for the list to include all of those act properties *F* such that there is a character trait consisting in proneness to performing *F*-ish acts, and it is a *virtue*. Thus, being just is on the list since there is a character trait consisting in proneness to performing just acts—I am taking the noun 'justice' to refer to it—and it is a virtue. For similar reasons, being generous, being kind, and being considerate are also on the list. In light of that condition on membership on the list, I shall call these properties *virtue properties*.

III

The story I am telling says that the virtue properties are second-order ways of being good, resting on the first-order goodness-for.

How resting? On subjectivist views, for an act to possess a virtue property, it is only intention that matters, and not success. On those views, therefore, an act can possess a virtue property and be good for no one at all, since the world may conspire against the agent and make his just or generous or kind act misfire. I think that a mistake. My own, objectivist, view is that, if a man intends to be acting justly or generously or kindly, that may (or may not) speak well for him personally, but his intention in acting does not at all settle the question whether he has *in fact* acted justly or generously or kindly. I shall not argue the matter here, however.

---

[8] It has to be conceded that from this point of view, being just is not as simple a case as the others on the list. Being generous and being kind are possessed only by acts and people. Being just has a wider range of application: some states of affairs and forms of government are also just. Are *they* just only derivatively? And in what sense are they if they are? One could instead argue that it is states of affairs that are (nonderivatively) just, and everything else (acts, people, governments) at best derivatively just. I must leave these things open. Fortunately, nothing in my story relies on any particular decision about them.

For even if it is success (and not mere intention) that matters, it is plain enough that an act can possess a virtue property and not be *on balance* good for people, either because it is good for fewer than an available alternative or because it is less good for those affected than an available alternative. The example of Alfred, who is under threat by the Mafia, is of the former kind, that is, good for fewer. The example went like this: if Alfred does not kill Bert, the Mafia will kill three other people. Those who offer this as a countercase to Moore's story believe that Alfred's killing Bert would be unjust; they believe that Alfred's refusing to kill Bert would be just, and that is surely right. Now, Alfred's refusing to kill Bert would be good for Bert. But it would be bad for the three others. So the fact of the refusal's being just is entirely compatible with the refusal's being good for fewer than the available alternative, and thus with its failing to be on balance good for people.

So we cannot say that an act possesses a virtue property only if it is on balance good for people. If the virtue properties rest on goodness-for, then they do so in a more complex way than that.

I cannot say now exactly what that more complex way is. There is a gap in my story here. Still, there is a familiar idea about the virtues that I think must surely be right, and if it is, then the following may anyway point toward a way of filling the gap.

What I have in mind is the idea that the fact of there being people who possess the virtues is good for us. On some views something stronger is the case, namely, that a person's possessing a virtue is good for that very person. Perhaps that stronger idea is correct. Even if not, however, there being people who possess the virtues is plainly good for the rest of us.

Justice is an obvious case. Justice is proneness to performing just acts; that is, it is proneness to doing what one *owes* to others—whether in the way of keeping one's word, or refraining from taking advantage of the weakness or ignorance of others, or carrying one's fair share of the community's burdens, and so on. No doubt a particular just act may not be on balance good for people (as in the example of Alfred). But it is better for us that the people among whom we live be just than that they not be just. Indeed, this is not merely better for us, but essential to us, since we can form a community at all—and thereby obtain benefits which are essential to us and which only community can provide—only if a substantial number of those among whom we live are just.

I stress: whatever else may be true of the people among whom we live, it is better for us that they be just than that they not be just. Even if they are not generous, not kind, and not considerate. For it would be worse for us if they were also not just.

What I suggest now is that we should take this to be, not merely a fact about justice, but what marks it as a virtue.

Is honesty a virtue? Well, what exactly is it? If honesty is (as some people think it is) a sheer unwillingness to lie, come what may, then on my view it should not turn out to be a virtue. (Indeed, on my view it is a peculiarly unattractive form of self-righteousness, and thus a minor vice.) And it does not turn out to be a virtue on this account of the matter, for it is better for us if the people among whom we live are ready to lie when they do not owe it to their hearers not to lie and lives will otherwise be lost. If honesty is, more narrowly, an unwillingness to lie where one does owe it to one's hearers not to lie, then it is a virtue, a subvirtue falling under justice.

Is prudence a virtue? Again: it depends on what prudence is. If prudence is (as some people think it is) entirely self-advantaging, if, that is, it is carefulness only where one's acts may cause harm to oneself, then prudence should not turn out to be a virtue, and does not on this account of the matter. If prudence includes, more broadly, carefulness where one's acts may cause harm to others, then it should turn out to be a virtue. But then so understood, it, too, is a subvirtue falling under justice.

For obvious reasons, we can call this nested cluster of virtues the *reliance virtues.*

There is a second cluster of virtues, which we can call the *virtues of concern.* They include generosity, kindness, and considerateness, and they have two features, the first of which marks them as virtues, the second as virtues distinct from the reliance virtues. The first is that they are pronenesses to doing what is good for others at a cost, at most, to their possessors. (It is not generosity in me to take from you to give to others. On the other hand, I say 'at most' since a kind act may be entirely costless; by contrast, an act is not generous unless it is in a measure costly to its agent.) In light of their having this first feature, they have a direct connection with goodness-for, more direct than the reliance virtues. It is plain that whatever else may be true of the people among whom we live, it is better for us that they possess these traits than that they do not.

Their second feature is that the acts these virtues are pronenesses to performing are not called for by justice; that is, it is not unjust to omit them. (However much good it may do my grocer for me to pay his bill, paying it is not an exercise of generosity or kindness or considerateness.) In light of their having this second feature, they are not among the reliance virtues: a person can be just without possessing any of them.

A person might have a proneness to "maximizing goodness-for"—that is, he is prone to doing a thing wherever it would be on balance better for people that he do it than that he not do it. That proneness also has a direct connection with goodness-for. But two things should be noticed about it. On the one hand, it is not a virtue of concern, since it fails to meet the first of the two conditions I mentioned, for when I do what I owe it to you not to do, because it would be on balance better for others that I do it, then my act imposes a cost (not on me, but) on you. On the other hand, while it might have seemed, off the cuff, to be better for us that those among whom we live possess this trait than that they do not, that is not true. If we cannot count on others' keeping their word to us unless we are so far lucky as that their keeping their word to us maximizes goodness-for, then we (simply) cannot count on others' keeping their word to us. The bearing of this on the possibility of our forming a community is obvious.

Perhaps the virtues of concern nest in the way in which the reliance virtues do, with generosity dominating them as justice dominates the reliance virtues. No matter, for present purposes.

What about courage? Courage appears on many philosophers' lists of the virtues, but I have so far postponed discussion of it. Suppose we believe that (i) courage is a virtue. What is it? Suppose we believe, as I do, that (ii) courage is (roughly) steadfastness of purpose, proneness to "standing firm," even in face of danger. This says nothing about what the courageous person's purposes are—it leaves open what they are—and that may seem to make trouble for my test for being a virtue, which says that (iii) a trait is a virtue just in case whatever else may be true of the people among whom we live, it is better for us that they have the trait than that they not have it. For if the people among whom we live are just, then all is well, it is better for us that they also be courageous. But if they are unjust, then it is (much!) better for us that they be cowards. In sum, if (ii) and (iii) are true, then courage is not a virtue. But (i) says it is a virtue, so something has to give.[9]

---

[9] If we accept (ii) we are committed to accepting that it is entirely possible for a person to be both courageous and a villain. Geach and Foot reject this possibility; see Geach, *The Virtues* (New York: Cambridge, 1977); and Foot, "Virtues and Vices," in her *Virtues and Vices* (Cambridge: Blackwell, 1978). But I have been unable to become clear about what alternative construal(s) of courage they would have us adopt.

I suggest that it is (i) that should give. Courage subserves justice in the sense that just acts require at least a minimum measure of courage. (It also subserves generosity, for generous acts may also require courage, and very generous acts may require great courage.) But it also subserves injustice, for unjust acts also require at least a minimum measure of courage, and very unjust acts may require great courage. It is hard to see how a trait that subserves a major vice can plausibly be thought of as, itself, a virtue.

Courage is not unique in this respect. Compare industriousness. (What is this industrious man industrious *at*?) A minimum measure of industriousness is required by just acts; but very unjust acts may require it, too. Again, compare loyalty. (Who or what is this loyal man loyal *to*?) A minimum measure of loyalty is required by just acts; but very unjust acts performed by groups are likely to require it, too. Similarly for conscientiousness. (What are the contents of this conscientious man's conscience?)

In sum, then, I suggest that we should not include courage—or industriousness, loyalty, or conscientiousness—among the virtues.

Let us return now to what I was calling *virtue properties*. I said I meant for the list to include all those act properties *F* such that there is a character trait consisting in proneness to performing *F*-ish acts, and it is a virtue. Justice and generosity are pronenesses to performing just and generous acts, and they are virtues; so being just and being generous are virtue properties. Courage is not a virtue; hence being courageous is not a virtue property.

That, I think, is as it should be. The properties I am calling virtue properties are those I indicated at the beginning of section II: being just, being generous, being kind, being considerate, and so on. I said that an ascription of these is praise, and that it seems clear that these are further ways of being good—that is, ways of being good in addition to the first-order ways of being good. Is an ascription of being courageous praise? When we ascribe being courageous we typically are praising; that is because we do not typically ascribe being courageous where we think ill of the act in question on other grounds. (Nobody says of a particularly villainous act, "Well, it was good in one way anyway: it was courageous." Nobody for a moment considers giving medals for courage to courageous villains.) These facts can be explained if we take seriously that an act's being courageous just is its having been performed despite danger to its agent, for this suggests that the praise goes to the act for occurring at all. Suppose a man risked his life to save another. That was a very generous act. It was also very courageous. To praise it on the former ground is to praise it for what it was; to praise it on the latter ground is to praise its occurring despite the danger to its agent. That is why praise of an act on ground of its being courageous is parasitic on there being other grounds for welcoming the act. (It also explains why no praise goes to courageous acts of villainy.)

All of this is rough, however, and I must leave it open. As I said, there is a gap in my story here. I have wished merely to point toward what *may* be a way of filling it.

IV.

What we have so far is this. The first-order ways of being good are the likes of being good for use in, good at, good for, and so on. There are also second-order ways of being good: among others, the virtue properties being just, generous, kind, and so on. These rest on a particular first-order way of being good, namely, goodness-for. (Exactly how do they rest on goodness-for? I made a suggestion, but it was no more than that.)

If what we have so far is correct, then we are home free: the rest of the story is briefly told, for we can get to the right relatively easily from the virtue properties. In short, what we are morally required to do is to avoid their contraries. Morality requires us to *do* a thing if and only if not doing it would be unjust, or mean, or cruel, and so on. Morality requires us *not* to do a thing if and only if doing it would be unjust, or mean, or cruel, and so on.[10]

Some philosophers have been saying in recent years that moral philosophers should look at the virtues, and some among them have said, more strongly, that moral philosophers should not only look at the virtues, they should also overcome their fixation with moral requirement. I agree that it would be profitable to look at the virtues. But that is because of the importance of their contraries, the vices, to supplying an account of what morality requires us to do.

My way of putting the matter might seem more complicated than it need have been. Why put it in the negative rather than the positive? Why say we are morally required to avoid injustice, meanness, and cruelty instead of saying that we are morally required to act justly, generously, and kindly? Moral requirement seems to me to be a rather weak notion: the standard it imposes is that of mere human decency. Now, generosity, kindness, indeed justice itself, can be very costly, and there are cases in which paying those costs would be supererogatory. Morality may require us to pay those costs, but I suggest that it does so only where refusing to pay them would be mean, cruel, or unjust.

---

[10] For my own part, I believe we should accept an objectivist view of the requirements of morality, that is, a view according to which a person's having done what morality requires him to do turns on success rather than on intentions—just as, for my own part, I believe we should accept an objectivist view of the virtue properties (and their contraries), that is, a view according to which an act's possessing a virtue property (or its contrary) turns on success rather than on intentions. (See the second paragraph of section III above. I regret having wobbled about this matter in *Moral Relativism and Moral Objectivity*.) I do not argue for these things here, however. What is surely plain, and is in any case required by this part of my story, is that we should be either objectivist about both or subjectivist about both.

To return just briefly, then, to Alfred and the Mafia: I said it is hard to see what could make Alfred's killing Bert wrong if it is not bad, and then hard to see how the Mafia's killing the other three people could fail to be three times as bad. Now we can say: the way in which Alfred's killing Bert would be bad is that it would be unjust. The way in which the Mafia's killing the three others would be bad is that it, too, would be unjust. Morality, therefore, requires both Alfred and the Mafia to refrain. That they will not refrain from killing their three if he refrains from killing his one leaves that conclusion entirely unaffected.

This solution to the problem relies on one's willingness to agree that the fact that the Mafia will not refrain from killing their three if Alfred refrains from killing his one does not mean that Alfred owes it to the three to save them from the Mafia. It seems to me plain that that fact does not mean this. Fortunately for my story, there is reason to think that Alfred's not owing it to the three to save them can be explained by appeal to the first-order goodness-for. A moral regime under which I can make you owe people something simply by threatening not to pay my own debts to them is a moral regime with a massive free-rider problem, and thus is a regime that would be bad for all of us.

Two points remain to be taken note of. First, a potential difficulty for my account of moral requirement emerges as follows. Suppose we believe that the following is possible:

(1) Bert's not doing such and such would be unjust, but Bert's doing the such and such would be cruel.

Then my account of moral requirement yields that the following is possible:

(2) Morality requires Bert to do such and such and morality requires Bert not to do it.

Is that possible?

On some views, (2) is entirely possible.[11] If it is, then the potential difficulty is not an actual difficulty.

---

[11] See the literature on "moral dilemmas"—for example, ***Moral Dilemmas and Moral Theory***, H. E. Mason, ed. (New York: Oxford, 1996).

But suppose we believe (as I do) that (2) is not possible. Then we may seem to have an actual difficulty, for either we must give up the account of moral requirement that I have offered, or we must agree that (1) is not possible. Which to do?

It will not surprise that I suggest we should agree that (1) is not possible. I have no argument to the effect that it is not. But I simply cannot imagine how the following could be true: justice in a given case calls for doing what it is cruel to do. Justice may, of course, require harming people, as, for example, where it calls for imposing a severe punishment; proceeding, where it does, is not cruelty. (What is cruel is hurting or harming gratuitously.) Similarly, I believe, for any other pair of contraries of virtue properties. So for my part, too, the potential difficulty here is not an actual difficulty. But others may think otherwise.

The second point that remains to be taken note of is that my account of moral requirement construes it as entirely other-regarding. That is a consequence of the fact that my account of the virtues, and therefore my procedure for picking out the virtue properties, construes them as fundamentally social: that is, my account of what fixes whether a character trait is a virtue, is its effect on others. But that a story about the right and the good yields that moral requirement is other-regarding seems to me exactly as it should be. Gluttony, fecklessness about one's own interests, and excessive timidity are a bad business; they are certainly bad for their possessor, but on my view they are no breach of moral requirement—unless they issue in injustice, meanness, cruelty, and the like. That is because, as I said, moral requirement seems to me to be a rather weak notion. I said that the standard it imposes is that of mere human decency, and what I meant was human decency in dealings with others.

V

What precedes is two thirds of the story about the right and the good I find attractive. In sum, I began with the first-order ways of being good, and went up from there to the second-order ways of being good that consist in the virtue properties, and then up from their contraries to the requirements of morality. We need now to return to the first-order ways of being good, and go down from there.

What I mean becomes clear if we return to Moore's story for a moment. I said that Moore's story begins with the good. Some things are good, Moore said, and some things are not; so there is such a property as goodness. I said that Moore had much to say about which things do and which things do not possess the property goodness, but that I would not summarize that part of his story. My reason should now be clear: there is no such property as goodness. Now, anyone who likes the story I have been telling so far is going to need to say something about which things are good in the first-order ways. What makes a thing be good to look at? What makes a person be good at hanging wallpaper? What makes a thing be good for use in making cheesecake? In one respect, these questions seem even more pressing than their analogue in the case of Moore's story, for surely the first-order ways of being good are not a mere happenstance clutter. There must be something they have in common that marks them all as ways of being *good.*

Expressed in the formal mode: it is not mere happenstance that the word 'good' appears in the expressions 'good to look at', 'good at hanging wallpaper', 'good for use in making cheesecake', and so on: its meaning makes a contribution to their meanings, and we need to see what that contribution is.

Intuitively, for a thing *X* to be good in one of the first-order ways is for *X* to *benefit* someone or some thing *Y* (which might or might not be *X* itself) in the appropriate way, or to be capable of doing so. There has to be something "in it" for some *Y*, a gain or potential gain of the appropriate kind, if *X* is to be good in a first-order way. That, my story says, is what the first-order ways of being good have in common.

What are the ways in which *X* might benefit *Y*? *X* might please *Y*. *X* might answer to *Y*'s wants. There are other possibilities, too, one of which will emerge as important in the following section.

Take being good to look at, taste, or listen to. A thing that has one or other of those properties has them in that it pleases people: looking at it pleases, tasting it pleases, listening to it pleases, and so on.

Take being good at doing something, as it might be, hanging wallpaper or playing chess. A person who is good at hanging wallpaper is one who is capable of hanging wallpaper as people typically want that done when they want it done. (There is nothing in the nature of wallpaper that fixes how it should be hung; there is only how people typically want it hung.) A person who is good at playing chess is one who is capable of winning when playing with experienced players, thus is capable of answering to his own wants when playing with them. (Unlike wallpaper hanging, chess has rules that fix what counts as winning, and you are not playing chess on a given occasion—you are only dabbling, playing *at* chess—if you are not then trying to win.)

Take being good for use in doing this or that. Spring-form pans are good for use in making cheesecake: that is because getting a cheesecake out of an ordinary pan is hard, and results in a mangled cake, and people who want to make cheesecake typically want the making of it to be easy, and the result to be smooth and round.

Moreover, where what is wanted varies from place to place, so does what is good for use in doing this or that. In Norway, lye is good for use in preparing codfish; that is a function of Norwegian wants—not merely of what pleases them but also of what is and is not cheaply available there. Not so in America. (It would be weird to say that lye is really good for use in preparing codfish everywhere, it is just that Americans do not know that. It would be equally weird to say that lye is not good for use in preparing codfish anywhere, it is just that Norwegians do not know that.)

I have so far mentioned three of the first-order ways of being good: first, being good to look at, taste, listen to; second, being good at doing a thing; third, being good for use in doing a thing.[12] Let us focus on a fourth, namely, being good for a thing. That way of being good is crucial to my story since it is what the story says the virtue properties rest on.

VI

Very many inanimate objects *Y* are such that for some *X*, *X* is or would be good for *Y*. I begin with artifacts. Regular oiling is good for a lawn mower. How so?

All artifacts have one or more of what might be called *design functions*, where 'Among the design functions of *Y* is to *A*' is true just in case *A*-ing is among the things *Y* is designed to do.[13]

Where *Y* is an artifact, *Y* is designed to *A* in that it was designed by a human designer to *A*. That is, a human designer selected some features *F* for *Y* because *Y*'s having *F* would increase the likelihood that it would *A*. Thus, among the design functions of any lawn mower is to enable its user to mow lawns with it, since some of its features were selected for it because its having them would increase the likelihood of its enabling its user to mow lawns with it. Another of the design functions of a lawn mower is likely to be to enable its user to mow lawns safely with it, since some of its features are likely to have been selected for it because its having them would increase the likelihood of its enabling its user to mow lawns safely. (I doubt that any artifacts have only one design function. Whatever the use to which a manufactured item is to be put, some of its features are likely to have been selected, not just to increase the likelihood of its being usable for that purpose, but also to increase the likelihood of its being safe to use for that purpose.)

---

[12] For more detail, see *Moral Relativism and Moral Objectivity*. I should add, however, that the account of goodness-for which appears in the following sections is different from the one I offered there.

[13] I adapt this notion *design function* from what is currently the most widely accepted analysis of functions; I have been particularly helped by Philip Kitcher, "Function and Design," *Midwest Studies in Philosophy*, XVIII (1993): 379-97. (Design functions are different from what I shall later call *use functions* and *role functions*.)

Now, I said in the preceding section that intuitively, for a thing *X* to be good in one of the first-order ways is for *X* to benefit someone or some thing *Y* in the appropriate way, or to be capable of doing so. I suggest, then, that we should say: if *Y* is an artifact, then (i) *X* is good for *Y* just in case *X* benefits *Y*, and (ii) *X* benefits *Y* by and only by conducing to *Y*'s doing what it is among *Y*'s design functions to do. For example, regular oiling is good for a lawn mower and that is a consequence of the fact that regular oiling benefits it, which it does by conducing to the lawn mower's doing what it is among its design functions to do, namely, to enable its user to mow lawns with it.

What is important is that on this account of the matter, the benefit supplied by *X* does not consist in anybody's being pleased, or in anybody's wants being answered to: the benefit supplied by *X* is to the artifact *Y* itself. That seems to me as it should be. Regular oiling is good for a lawn mower, and that is not because regular oiling of a lawn mower pleases anybody, or answers to anybody's wants. No doubt it was the wants of consumers, and therefore the wants of manufacturers, that fixed that lawn mowers were designed in such a way as to have, among other design functions, the design function of enabling their users to mow lawns. No doubt people buy lawn mowers nowadays because they want to be able to mow their lawns, and are pleased when they are able to do so. But regular oiling would still be good for a lawn mower even if people stopped wanting to mow their lawns, and even if manufacturers therefore stopped making lawn mowers. Suppose that starting next year, lawn mowers pile up in garages and warehouses, rusting away from lack of care because of lack of users. Oiling a lawn mower would not benefit any person, but would all the same benefit *the lawn mower*; and oiling it would benefit it precisely because, though nobody wants to use it to mow lawns, oiling it would conduce to its being in a condition in which it *could* be used to mow lawns. That, after all, was what it was designed to do.

We can reput the matter in the following way. Many things have one or more of what might be called *use functions*, where 'Among the use functions of *Y* is to *A*' is true just in case some people do in fact use *Y* to *A*. I should think that as things are, the use function of all (working) lawn mowers is to mow lawns, since some people do use them to mow lawns. But two facts remain. (a) The design functions of a lawn mower remain what they are no matter what its use functions come to be. Thus, the design functions of a lawn mower continue to include 'to enable its user to mow lawns with it', even if the lawn mower—indeed, even if all existing lawn mowers—cease to have any use function (as by being allowed to rust in warehouses) or acquire a novel use function (as by coming to be used merely as garden ornaments). (b) Where a lawn mower's use functions diverge from its design functions, what is good for it is not conducing to its doing what it is among its use functions to do, but rather conducing to its doing what it is among its design functions to do.

Similar points hold about some animate things, which I shall turn to in the following section.

Let us first look briefly at inanimate objects that are not artifacts. If *Y* is inanimate but not an artifact, *Y* has no design functions. Yet it is possible that something, *X*, is or would be good for it. It is plausible to think that, if *X* is good for such a *Y*, then that is because *X* conduces to *Y*'s being in a condition that answers to people's wants. Where people have conflicting wants in respect of *Y*'s condition, it may (or may not) be easy to say whose wants count. Imposing regulations on disposal of wastes in the Charles River in Massachusetts was good for it. How so? It conduced to the river's being in a condition that answered to the wants of those who wanted to fish in it, swim in it, sail small boats on it, and so on. Why is it their wants that fixed what was good for the river? Why not the wants of those who wanted to dump wastes in it? I am going to bypass this question. For present purposes, what matters is only that it is people's wants that fix what is good for it. If the condition of the Charles stopped mattering to people—and were never again going to matter to anybody—then nothing at all would any longer be good (or bad) for it. Rivers are like wallpaper and unlike lawn mowers: just as there is nothing in the nature of wallpaper that fixes how it should be hung, so also there is nothing in the nature of rivers that fixes what condition they should be in or come to be in.

VII

All animate objects *Y* are such that for some *X*, *X* is or would be good for *Y*. Let us begin with plants. Watering a certain plant might be good for it. How so?[14]

Plants, too, are designed to do things, and thus they, too, have design functions. Where *Y* is a plant, *Y* is designed to *A* if it was designed by nature to *A*. That is so if *Y* has some features *F* because possession of *F* by *Y*'s ancestors increased the likelihood that they would *A*, where increasing the likelihood that they would *A* conduced to their reproductive success.[15] Thus, among *Y*'s design functions might be to attract pollinators of kind *K*; that is so if *Y* (as it might be) flowers at a certain time of the year because its ancestors' flowering at that time of year increased the likelihood that they would attract pollinators of kind *K*, and doing so conduced to their reproductive success. Another of *Y*'s design functions might be to protect itself from disease; that is so if *Y* (as it might be) grows a scab or scar when injured because its ancestors' growing a scab or scar when injured increased the likelihood that they would protect themselves from disease, and doing so conduced to their reproductive success.

Indeed, I should think that every animate object *Y* has the following design function: to survive. (More precisely, no doubt: to survive to a time, if any, after which further survival would impede reproductive success.) For every animate *Y* has some features *F* because its ancestors' possession of *F* increased the likelihood that they would survive, and doing so conduced to their reproductive success.

---

[14] It has been suggested to me that the answer is simple: watering the plant would be good for it just in case, and because, it *needs* water. And, quite generally (and in particular, for artifacts as well as plants), that

(1) Getting alpha would be good for *Y*.

is analyzable into

(2) *Y* needs to get alpha.

This will not do, for (1) and (2) are not even equivalent. What is equivalent to (2) is surely:

(3) Not getting alpha would be bad for *Y*.

and (1) is not equivalent to (3). (For very many, if not all, things *Y*, there are things alpha such that getting alpha would be good for *Y* though not getting alpha would not be bad for *Y*—and those are therefore things which *Y* does not really need.)

It may be suggested that we should say, "Well, anyway, (3) is analyzable into (2)." Why say that rather than that (2) is analyzable into (3)? In fact, if we are unclear about what makes one of them true, we are not helped by being told that what makes it true is what makes the other true—any source of unclarity about one is equally a source of unclarity about the other.

[15] "[O]ne of Darwin's important discoveries is that we can think of design without a designer"—Kitcher, p. 380. Many people, including Kitcher, recommend taking *Y*'s recent ancestors to be what matter. I bypass all the important questions that arise here.

I suggest, then, that we should say about plants what I suggested we should say about artifacts: if *Y* is a plant, then (i) *X* is good for *Y* just in case *X* benefits *Y*, and (ii) *X* benefits *Y* by and only by conducing to *Y*'s doing what it is among *Y*'s design functions to do. In particular, watering a certain plant is good for it if watering it conduces to its surviving.[16]

So just as in the case of artifacts, what fixes what is good for a plant is not what pleases anybody, or what answers to anybody's wants. Feeding the dandelions in my front lawn a high nitrogen fertilizer would be good for them even though what *I* want is that they die, before reproducing if possible. A plant's use functions, if it has use functions, are irrelevant, except in so far as their use functions converge with their design functions.

VIII

People are obviously a more complicated affair than plants. They, too, have design functions: there are many things they are designed to do since they were designed by nature to do them. Among their design functions is to protect themselves from disease: like plants, they grow a scab or scar when injured because their ancestors' doing this increased the likelihood that they would protect themselves from disease, and doing so conduced to their reproductive success. Another of their design functions is to survive.

---

[16] Some people say that what conduces to a plant's survival is good for it, but that what conduces to a plant's reproducing is good only for its species, and not for the plant itself. Why should we believe this? (Can those who say it harbor the idea that, like many human beings, plants want to survive but do not care about reproducing?) A further difficulty is that it is by no means clear what could count as something's being good for a species. Is it bad for a species to become extinct? No doubt it might be bad for us that a species become extinct: What could possibly make it bad for the species itself to do so?

I am grateful to Paul Horwich for drawing my attention to the fact that many people say that a species' becoming extinct would be bad—not bad for us, not bad for it, but just plain a *bad thing*. I have said that there are no such properties as goodness and badness; so what can those people be taken to mean? We should note, first, that nobody says this about just any species. I gather that hundreds of species of ants (or is it termites?) become extinct every day: Is anyone seriously inclined to call that just plain a bad thing? Pandas are another matter, however. Dear living teddy bears! But that points to the way in which their becoming extinct would be bad: we would lose something if they became extinct, and thus their becoming extinct would be bad for us. (There are other losses besides losses of the useful. Consider losing your old childhood nonliving teddy bear. Or your child.) So far as I can see, it is in being bad for us, and only in being bad for us, that their becoming extinct would be bad.

The species "human being" is yet another matter. If humans become extinct because of something that causes the last of them pain, then that would be bad for them. Suppose they become extinct because the last of them want not to reproduce, and that that want is "ideal" in a sense to emerge in the text below. (Not due to ignorance, and so on.) Thereafter the world would go rolling on to its end without any human intelligence in it, without any high intelligence in it at all unless some other highly intelligent species arises. Would that be bad? (In what way?) Is the sadness that prospect provokes in some people like the sadness that some people feel at the prospect of their dying without issue? (*Is* dying without issue bad? In what way?) I leave these dark questions open.

But there are also things they are designed to do that they were not designed by nature to do, but rather by themselves. Just as a human designer can select some features *F* for an artifact because the artifact's having *F* will increase the likelihood that it will *A*, so also can human beings select some features *F* for themselves because their having *F* will increase the likelihood that they themselves will *A*. For example, a man may select unusually strong muscles for himself (going into training to acquire them) because his having them will increase the likelihood that he will win at weight lifting.

Some people are happy to accept the following sufficient condition for use of the term 'goals' (and analogous conditions for 'aims' and 'purposes'):

(1) *Y* has among its goals to *A* if *Y* has among its design functions to *A*.

It is a consequence of (1) that plants and lawn mowers as well as people have goals; the people I refer to are happy to accept that consequence. I am sure that at least some among them also accept:

(2) *Y* has among its goals to *A* if *Y* chooses behaviors it thinks will increase the likelihood of its *A*-ing, and does so because it wants to *A*.

It is not a consequence of (2) that plants and lawn mowers have goals. But some things meet both sufficient conditions. My weight lifter has among his design functions to win at weight lifting, and so has that as his goal under (1); he also chooses behaviors he thinks will increase the likelihood of his winning at weight lifting, and does so because that is what he wants to do, and so has that as his goal under (2).

On the other side are people who reject (1), and opt for a strengthening of (2), namely:

(3) *Y* has among its goals to *A* *if and only if* *Y* chooses behaviors it thinks will increase the likelihood of its *A*-ing, and does so because it wants to *A*.

According to (3), only people and higher animals have goals.

There is no need for us to decide which side is correct. I shall accept (3), but largely because doing so provides a perspicuous way of abbreviating the theory I think we should accept.

It is a familiar theory. It says, in short: if *Y* is an adult person, say Alfred, then (i) *X* is good for Alfred just in case *X* benefits Alfred, and (ii) *X* benefits Alfred by and only by conducing to his reaching one or more of his goals. (It may pay to make explicit that I mean to include enabling him to reach his goals under 'conducing to' his reaching his goals.)

But that is too short, for not just any goal Alfred happens to have counts for the purposes of this theory: a goal counts only if it meets two kinds of constraints: (a) Alfred does not have the goal because of ignorance of, or lack of careful attention to, the costs of doing *A*, including opportunity costs, and (b) Alfred does not have the goal because of such improper "preference bendings" as threats, drugs, or a grossly unjust political regime.

A familiar way of accommodating these constraints is by reputting the matter counterfactually: what counts is what Alfred would aim at in ideal conditions of full information about costs, assessed "in a cool hour," and lack of improper preference bendings.

Why do those constraints have to be imposed? Alternatively put, why does it matter that Alfred would aim at this or that in ideal conditions, given he does not in fact aim at the this or that? The answer is that what is in question here is what *really* benefits Alfred, not what he may happen to think would benefit him—off the cuff, in ignorance, under threat, and so on.

In sum, then, *X* is good for Alfred just in case *X* benefits him in that it conduces to his reaching one or more of (what I shall call) his *ideal goals*.

I said that this is a familiar theory, but it is certainly not universally accepted. Might it not turn out that Alfred's ideal goals are such that reaching them precludes his having family and friends, or a wide variety of intellectual and aesthetic enjoyments, or even mere good health? Would it not benefit him to conduce to his life's including those things, despite his ideal goals? Well, would it? Proponents of the familiar theory reply: if we think it would, then that is because having those things in our own lives is important to us, but it is mere hubris to suppose that getting them benefits every person, whatever his ideal goals may be, and however drastically making room for them in his life might interfere with his chances of reaching his own ideal goals. Who are we to tell the saint or research scientist or great chess master to make some friends, read some good novels, listen to some good music, keep out of unhealthy environments—assuming that his actual goals are his ideal goals?

What if Alfred's ideal goals include or require doing harm to others? Proponents of the familiar theory reply that while it *may* be true that what benefits one does no harm to others, this is by no means an a priori truth.

These replies seem to me to be right. So the familiar theory also seems to me to be right.

Some things that follow from that theory might be worth drawing attention to. In the first place, Alfred's use functions, if he has any, are irrelevant to what is good for him; what is good for him is not fixed by what benefits others who use him for this or that purpose but by what benefits him.

Perhaps people do not on the whole have use functions. But they mostly (all?) do have what might be called *role functions*. If Alfred's job is to keep the books at the local shoestore, then keeping those books is among his role functions. Fine. But while conducing to his keeping those books may be good for him, it also may not be.

A third and more interesting consequence of the familiar theory is this: not only do use functions and role functions fail to fix what is good for a person, but so also do design functions. Suppose that Alfred was designed by nature to *A*. Conducing to his *A*-ing may conduce to his reaching one or more of his ideal goals. But it may not. Doing so may even conflict with his reaching one or more of his ideal goals. If conducing to his *A*-ing does not conduce to his reaching an ideal goal, then the familiar theory tells us that doing so is not good for him. So design functions have not the role in fixing what is good for a person that they have in fixing what is good for artifacts and plants.

But this, too, seems to me to be right. It is among Alfred's (naturally selected) design functions to survive. But the hero or saint may try to save the life of another at risk to his own life. It cannot at all plausibly be thought to be good for him to prevent him from doing so.

Again, it is among Alfred's (naturally selected) design functions to be able to engage in such physical activities as running. It is hard to imagine a person whose ideal goals are so eccentric that an ability to run would not in any way help him to reach them. But suppose that is true of Alfred: suppose he cares not the least about being able to run. Then conducing to his being able to do so is not good for him. That seems to me very plausible.

Things are otherwise when it comes to Alfred's body parts. It is among the (naturally selected) design functions of Alfred's legs to enable Alfred to run. Legs have no goals, however, so what is good for them is fixed by what conduces to their doing what it is among their design functions to do. Alfred's ideal goals are conclusive when it comes to what is good for him; they are irrelevant to what is good for his legs, just as they are irrelevant to what is good for his lawn mower and his plants.

IX

In sum, the story I have been telling went as follows. I began by saying that there is no such property as goodness; there are rather being good for use in, being good at, being good to, being good for, and the like. I called these first-order ways of being good. The second-order ways of being good include the virtue properties: being just, generous, courageous, kind, and so on. These, I suggested, rest on the first-order goodness-for. From them we can get to moral requirement by way of the following thesis: what we are morally required to do is to avoid their contraries. If these ideas are correct, then the right does rest on the good, though unfortunately not in the very simple way described in Moore's story.

I suggested also that the first-order ways of being good rest on benefitting in appropriate ways, which may involve pleasing, or answering to wants, or conducing to something's doing what it is its design function or ideal goal to do. This may well be the most controversial part of the story. It should be stressed, however, that the rest of the story is independent of this part. What the fact of this part's being unacceptable would mean (if it is unacceptable) is just that some better account of the expressions by which we attribute the first-order ways of being good had better be found. For it is not just happenstance that the word 'good' appears in all of those expressions: its meaning surely does contribute to their meanings, and we need an answer to the question how it does.

JUDITH JARVIS THOMSON

Massachusetts Institute of Technology

TEMPORAL HORIZONS OF JUSTICE

We are mortal creatures. Perhaps we might live human lives without seriously confronting death. But while we are alive, we cannot avoid making sense of life's meaning by locating ourselves in time. Some will tell elaborate narratives about their past, present, and future. Others will look on these hyperorganized accounts with disdain or despair, and content themselves with more fragmentary stories. But without any sense of past, present, future, we are nothing.

So much is obvious. But it is less obvious how our ongoing effort at temporal self-location shapes thinking about justice. I elaborate three temporal horizons that we are continually using to make sense of our lives. Each horizon frames time in a different way, irreducible to the others. Each is irreplaceable. We must make our peace with all three, despite their potential conflicts with the others. As we explore two of these horizons, we find ourselves on a path that leads to new insights into an old idea—that there is a basic difference between *corrective* and *distributive justice*. Exploring a third will lead us into less familiar, but equally fundamental, terrain: I shall call it the *sphere of relational justice*.

This tripartite view permits a fresh encounter with more familiar monistic and dualistic perspectives. Monists hope to work out a single idea of justice that will fit all kinds of disputes. Some have been notoriously simple-minded—as in Benthamite utility maximization or Posnerian wealth maximization. But there is nothing in monism that requires brutal simplicity. You can be as complicated as you like in developing your idea of justice—so long as you think that the right idea, once elaborated, should apply to all possible cases.

Dualists have no such expectation. They follow Aristotle and assert that justice comes in two varieties—corrective and distributive. John Rawls, for example, does not suppose that his famous difference principle should regulate our ideas of corrective justice. Robert Nozick[1] and many other conservatives have attacked Rawls on monistic grounds: If Rawlsian justice cannot be used to inform microdiscussions of particular cases, why should we trust it to regulate the basic structure of society?

An analogous dispute has been provoked by the rise of the law-and-economics movement over the last generation. This time, tradi-

---

[1] See *Anarchy, State and Utopia* (New York: Basic, 1974), pp. 204-13.

0022-362X/97/9406/299–317

tionalists tend to be dualists. On their view, lawyer-economists make a category mistake when they test the rules of contract, tort, and crime by one or another conception of efficiency. Instead, legal developments should be guided by the very different commands of corrective justice.[2] For lawyer-economists, this move merely mystifies potential irrationalities and inefficiencies in the name of an obscure label, signifying nothing.

Disagreements between monists and dualists do not break down on political lines: conservatives are not necessarily dualists, liberals are not necessarily monists, or the other way round. Despite the antiquity and intensity of the dispute, we do not seem to have much of an idea about what is dividing the two sides. Monists and dualists are generally content to accuse one another of moral blindness—for monists, dualists are perverse obscurantists; for dualists, monists are barbaric simplifiers—without trying to understand why their apparently intelligent opponents fail to appreciate the error of their ways.

I shall be suggesting that both sides tend to give priorities to different time frames, and that both are right in supposing their opponents miss something important. Once we understand the dispute in this way, we shall no longer be tempted to suppress the insights of either side. It may prove possible to incorporate the insights of both schools into a more complex, trinitarian, concept of justice.

I. CLOCK WATCHING

Let me begin the modern way—by inviting you to look at your watch. Unlike the denizens of previous civilizations, it is second nature for us to walk around with a device that constantly tells us that time is like a line in space. Just as we can subdivide a line into homogeneous and infinitely divisible units, we can subdivide the time line by glancing at our watch. The only difference is that once we have passed a point on the time line, there is no returning.

Now, you do not have to be Albert Einstein to know that this linear analogy between space and time has proved unsatisfactory for scientific purposes. But I am not doing science here, or even the philosophy of science. My aim is to elaborate the temporal frameworks used by ordinary men and women to make sense of the meaning of their lives. These frameworks provide the context within which these same ordinary people seek to realize justice in the law of a democratic society.

---

[2] See, for example, George Fletcher, "Fairness and Utility in Tort Theory," *Harvard Law Review*, LXXXV (January 1972): 537-73; Ernest Weinrib, "Legal Formalism: On the Immanent Rationality of Law," *Yale Law Journal*, XCVII (May 1988): 949-1016.

From this viewpoint, the most important event of the twentieth century is not the triumph of Einstein, but the triumph of the wristwatch. While modern men and women differ vastly in other respects, all are trained to measure out their lives in clock beats. It is one thing to arrive habitually late for appointments; quite another to refuse to wear a wristwatch or look at a clock. Such a blank stare would mark a person off as an alien presence: either she is a recent arrival from Mars or a philosophy student momentarily in the grip of John McTaggart's proof of time's unreality.

This essay does not try to vindicate clock time against its radical critics. I shall be struggling against a more common reaction. We sometimes take clock time so seriously that, despite Einstein, we suppose that it describes the nature of ultimate reality, that it is the only "real" way we can talk about time and its passage. But if we think about how we actually live out our lives, it is surprisingly easy to loosen the grip of the wristwatch upon our time consciousness.

I do not deny that the time line provides modern men and women with an irreplaceable temporal framework. But it is not the only form in which time, and its passage, appears to us. In contrast to clock time, these other forms are neither infinitely divisible nor homogeneous. They are nonetheless very real.

II. THE LIVED PRESENT

We live in the present. Each time we awake, we are thrown into the world of the now, and we cannot escape the task of interpreting the meaning of what is going on. Right now, for example, I am writing this line, and you are now reading it.

The present, for us, is not an infinitely small point on a line, but a structured experience. The key is recognition of change—whenever I notice that something significant has happened, this recognition serves to reframe my understanding of what is going on right now.

Insignificant things do not matter. As you read this essay, the significant event is the word or phrase. You are not aware, nor should you be, of the clock time it takes to decode a single letter—unless, of course, you reinterpret the nature of your structured activity and experience this essay as a challenge to identify the letters of the alphabet. Even then, your lived present would not be infinitely divisible into clock time. As you happily called out the letters on the page, you would be failing to notice many changes occurring at a different rate. While a clock can measure the passage of a nano-second, this is too short for a human being to focus his attention and begin to make sense of the world in which he finds himself.

So much for the lived present. A lot more can be said,[3] but despite my brevity, I am sure that you know what I am talking about. Rather than explore the now at greater length, I want to describe three ways of making sense of the now by placing it into longer, but differently constructed, temporal horizons. As I display each horizon, I shall also suggest how it shapes our sense of justice in a different way.

III. LIVED EXPERIENCE

First (and foremost?), each of us frames the now by placing it against a background of lived experience. I am now reading an essay, or playing baseball, or working on an assembly line. Each of these experiences lasts longer than the now, but none lasts too long. For one thing, their duration is limited by an internal understanding of the particular experience: if, say, I am playing baseball, and I strike out to end the ninth inning, I have come to the end of the game. While I may remember my experience a minute later as I brood about my strikeout in the shower room, I am perfectly aware that I am now "taking a shower"—a different experience altogether.

In contrast, when I went up to bat in the ninth inning, I was still experiencing the same game I started in the first inning—even if the game began several hours ago. It is the internal structure of my ongoing activity, not clock time, that marks off one lived experience from the next.

While an activity's internal structure is important, human physiology also imposes constraints. 'Let's stop for lunch' generally serves to call a halt to whatever we were doing; when we pick up afterward, we experience ourselves as starting up anew—the lunch break typically interrupts the sense of continuous engagement in a developing structure of experience which is the hallmark of the first temporal horizon. It is possible, of course, to defer termination of a lived experience by refusing to eat or sleep and by focusing attention on it alone. But soon enough, it will prove impossible to evade the demands of the body; and when you awake from your fitful slumbers, you will be opening your eyes to new lived experience(s)—even if they occur in the very same sequence as yesterday's.

The familiar rhythms of lived experience shape a first understanding of the demand for justice. Begin with the simple point that each lived experience is understood as having a structure: a set of expectations about the way in which one now can appropriately evolve into the next. If I am reading a philosophical essay, I may plan to read one paragraph and then the next; if I am working at McDonald's Restaurant, my expectations have a more complicated structure, allowing for more—but not infinitely more—appropriate ways in which my actions at one time may evolve into my experience at the next.

---

[3] My own thinking begins with the brilliant, if obscure, work of Edmund Husserl's *On the Phenomenology of Internal Time Consciousness*, John Barnett Brough, trans. (Dordrecht: Kluwer, 1991).

This elementary point suffices to formulate a basic question of justice: since each of your lived experiences has a structure, each is vulnerable to disruption. For example, you are now trying to make sense of this essay. Suppose Shmoe disrupts this effort by blasting a loud horn within earshot: What in the world, you say, is going on? Who does Shmoe think he is? What gives him the right to blast away at the structure of my lived experience?

Notice how naturally this temporal horizon frames the characteristic concerns of corrective justice. Your questions cannot be assimilated into some general complaint about the way society as a whole is organized. The problem is not society but Shmoe, and his disruptive behavior.

Not that Shmoe will find it impossible to answer your questions about his actions. Most fundamentally, he may *deny his responsibility* for the disintegration of your structured expectations—pointing, for example, to the fact that a third person, Blow, forced him to play his horn at gun point. In this case, you have misdirected your complaint, and had better redirect it appropriately.

Worse yet, you may find that there is nobody responsible for the disruption. After all, the world has not been organized to guarantee the integrity of all your lived experiences—you may simply be the victim of misfortune. Even if you can locate the responsible actor(s), he may plausibly respond with an excuse or justification for his conduct—perhaps he did not notice you sitting quietly in the shady corner, or perhaps he was sounding a fire alarm, or....

My aim is not to appraise any of these responses—this is the task of a full-blown theory of corrective justice. I am concerned with the temporal horizons that make these theories significant. From this point of view, the most important—and appealing—thing about corrective justice is its principled refusal to abandon the frame of lived experience, its insistence on understanding the problem of justice in terms that make sense within the narrow temporalities in which life is actually experienced from day to day. Rather than seek relief from disruption in the reorganization of social processes whose character only displays itself over larger tracts of time, corrective justice focuses remorselessly on the particular people who disrupt particular experiences at particular moments.[4]

Even within these confines, it focuses narrowly on those particularities which illuminate individual responsibility for the disruption. It generally casts a blind eye on whether Shmoe is rich or poor, but closely scrutinizes whether he noticed your presence in the room before putting the trumpet to his lips. It is the latter, not the former, that singles Shmoe out from your perspective of lived experience. To make this point, I shall say that the accent of justice within this time frame is on the *responsibility of persons*, with different approaches to corrective justice elaborating different variations on this theme.

For all the trendy talk about the "socially contingent" nature of human reality, it is hard to imagine how homo sapiens might dispense with the habits of experiencing life in chunklets of an hour or two, and of responding with outrage when others disrupt their patterned expectations about the way these chunklets "normally" evolve. It is this sense of outraged expectation which fuels the demand for corrective justice.

Indeed, corrective justice would be the only meaningful kind of justice if the structure of lived experience provided us with the only way we could ever understand the meaning of life. But it does not.

IV. RELATIONSHIPS

Particular experiences can be interesting, boring, or profoundly revealing. But they gain new meanings when we organize them into patterns that unfold over longer periods of time—patterns that I shall call *relationships*.

Imagine, for example, that I teach a particularly bad class—disorganized and obtuse, I lose touch with the concerns of my students. Retreating to my office in disarray, I place my recent pedagogic disaster into a larger pattern of activity: After all is said and done, am I a third-rate teacher? To answer, it is not enough to fixate upon my most recent lived experience. I must weigh the meaning of the other episodes that make up my past teaching activities: Remember how I managed to encourage that anxious but brilliant student to realize her potential? And there was the time that....

---

[4] While the search for the responsible actor begins with the disruption of a particular lived experience, it may not end there. Instead of blowing his horn in my presence, Shmoe may program his computer at Time$_1$ to emit a blast just as I am studying my book at Time$_T$. In such cases, actions at earlier moments become relevant—but only so far as a theory of corrective justice can link them up appropriately to the moment of disruption. It is this latter time frame which continues to serve as the decisive reference point. For a classic treatment, see H.L.A. Hart and Tony Honore, *Causation in the Law* (New York: Oxford, 1985, 2nd ed.).

Each element in this story may point to a particular lived experience; but the meaning and value I give to the activity point beyond particulars—to narratives that relate them into larger patterns of meaning. To put my point in a slogan: I *live out* my particular experiences but *I tell a story* about my relationships. In the story I tell myself, I may be a *good* teacher despite *bad* classroom experiences from time to time, or vice versa. Indeed, words like 'good' and 'bad' hardly suggest the moral complexities involved in the ongoing effort to assess the implications of particular experiences for the stories we tell ourselves about our ongoing relationships.

Most importantly, the value of particular experiences is not exhausted by the role they play in our stories that seek to place them into larger temporal patterns. Even if I am convinced I am a good teacher, this does not somehow cancel out the badness of particular experiences. Similarly, I may make a mess out of a relationship and yet recall certain shining experiences. While the two temporal horizons are complexly intertwined with one another, their value cannot be reduced to one another. But if this is true of one's personal life, is it also true about justice?

I shall explore the matter by considering the ongoing revolution of our gender morality. Certain aspects may be readily expressed within the horizon of lived experience. For example, if we all took seriously the idea that our sexual partners had the right to say 'no' to our advances, this important step would not require the definition of a distinctive time horizon. It simply redefines the Shmoe-like behavior that should count as disruptive of another's expectations about the structure of particular lived experiences.

But other aspects of the gender revolution can only be understood against the more extended horizon of meaningful relationship. Consider, for example, the typical situation of a woman who wishes to gain professional recognition as a first-rate lawyer or professor or car mechanic. Suppose that the men in her environment are not disruptive in any of the familiar abusive ways—they do not hurl sexual insults or demand sexual favors, much less assault or rape. They may even respond to her with labored politeness. To make things hard for ourselves, suppose that when payday comes around, our hypothetical woman worker is highly compensated, in full recognition of her contributions to the institution's marginal revenue product. She has only one problem: her fellow workers do not take her seriously. But what *precisely* does this mean?

Assume, for example, that a lady lawyer attends strategy session after strategy session in which the affairs of clients are discussed; on each occasion, she talks about as much as most participants, and nobody interrupts; but her contribution is never taken up by others in the subsequent give-and-take. Within the lived experience of each individual session, there is nothing odd about this fact. Everybody has been at meetings at which his or her contribution goes unnoticed—even if it actually has a good deal of merit. The problem arises only when our not-so-hypothetical lawyer puts the episodes into narrative form; only then does the pattern emerge that is the object of her grievance. These guys just are not taking her seriously!

Despite her hefty paycheck, she is a victim of injustice. If you agree, we cannot express this judgment within the temporal horizon of corrective justice. We require a new idea—that justice requires us to respect each others' right to tell stories of a certain kind about our ongoing relationships.

Having taken this step, two paths open before us. One is consequentialist: if her fellow workers began to take her contributions seriously over time, perhaps the woman would rise to new professional heights, and (after a transition) perhaps the law firm would find itself better off in terms of profits and camaraderie.

But perhaps not. It is, alas, perfectly possible that once her colleagues began taking her seriously, they would find lots to criticize about aspects of her performance which they had formerly disdained to notice; or perhaps the woman, having gained the professional recognition she had long sought, concludes in despair that "success" in the activity was not worth the long years of sacrifice it required. Nonetheless, even if the woman only gains misery upon winning professional recognition, the men were doing her an injustice in withholding it. Here, as elsewhere, the utilitarian account fails to express the categorical character of judgments of injustice: the wrong done to the woman goes deeper than the fact-specific balancing of costs and benefits allows us to appreciate.

Rather than guess at probable outcomes, we should look to the way the work environment suppresses a fundamental human capacity—the capacity to project the meaning of one's life over time by telling a story that includes meaningful participation in relationships. Without such a capacity, it would simply be impossible to live a human life. A social environment that denies a woman this possibility strikes at a crucial dimension of human dignity—one that is no less fundamental than the horizon of lived experience protected by a concern with corrective justice.

Just as corrective justice vindicates the very possibility of meaningful lived experience, so too relational justice vindicates the very possibility of meaningful participation in ongoing activities. If corrective justice requires each of us to act responsibly and avoid the wrongful disruption of lived experience, relational justice requires us to act responsibly by fairly recognizing the contributions others make to ongoing enterprises in which we participate. If we refuse to recognize this duty, we deny others their claim to dignity as fellow creatures in the search for a meaningful life.

Of course, a lot of questions must be resolved in the elaboration of this core idea:[8] Does it embrace the organization of the family, religion, sport, as well as work? In what ways should respect for the contributions of others be manifested? When, if ever, may participants legitimately waive their right to something I will call a *dignified* relationship?

But let us complete the view of the forest before glancing at a few trees. A third horizon awaits.

V. LIFE CHANCES

Think of yourself writing your autobiography. Surely, you would reflect on particular lived experiences that were decisive. You would also organize these, and many other, experiences into narratives that define larger patterns of enduring relationship. But your life story could not be reduced to any single relationship, let alone a particular experience. An adequate autobiography would reveal the intricate interplay of experiences and relationships that express the complex meanings of your life as a whole.

Within this life-time framework, a different question of justice takes center stage. Like all mortals, you have struggled for meaning in a world you did not create. You did not choose your parents, or your education, or the material and transactional resources with which you began life as an adult. Yet these resources have transparently shaped the life story you tell in your more autobiographical moments. If you had begun life with different resources, you would have had different experiences, engaged in different relationships. Have you, then, been unjustly handicapped—or advantaged—from the start?

This is the question that motivates liberal theories of distributive justice and accounts for certain characteristic features that can otherwise easily be parodied. The first is their high level of abstraction. Since we are scrutinizing the distribution of basic resources out of which people build the meaning of their lives, the analysis will avoid the examination of particular lived experiences or relationships—since, by liberal hypothesis, people should be free to organize them in the way they think best. Second, this temporal horizon also accounts for the theories' principled neutrality as to the nature of the good life: if basic resources are given out fairly, it is up to each of us to write our own autobiographies—for good or for ill. It is not for the liberal state to proclaim an authoritative answer to the mystery of life, and distribute resources on this basis.

---

[8] This idea lies at the foundation of Catharine MacKinnon's argument in her *Sexual Harassment of Working Women* (New Haven: Yale, 1979).

I deny, however, that this abstract and antiteleological characterization justifies the familiar charge that liberal distributive theories presuppose an "unencumbered self," whose dignity can somehow be detached from life's experiences and relationships. On the contrary, it is precisely because human dignity can only be achieved *through* experience and relationships that distributive justice seems a matter of crucial importance. Since our particular experiences and relationships are profoundly shaped by our initial claims to basic resources, those claims had better be allocated in a justifiable way. In focusing on the distribution of initial endowments, liberal theories do not presuppose an unencumbered self; they simply suppose an *autobiographical* self, a self that can look back over the shape of her life and recognize that its starting point and subsequent development was profoundly shaped by political decisions allocating economic and cultural resources.

Indeed, thoughtful people who *do* take a radically "unencumbered" view of the self and its dignity—whether that view be based on Stoic or Christian or Buddhist or Kantian sources—have never been at the forefront of the struggle for social justice. For these people, a concern with such matters is part of the problem rather than part of the solution: since it is (they claim) an illusion that our successes and failures in the world of experience really matter for the salvation of our souls, it really does not matter how much or how little we have been given in terms of basic resources. The only important thing is to seek a spiritual state of radical detachment from the things of this world. Within this framework, a passionate commitment to distributive justice is merely a snare and delusion, leading one to place undue importance on worldly matters that are obstacles to salvation.

But this is hardly the spirit of modern liberalism. Whatever my differences with communitarians, they are not to be found in a dispute on the metaphysical relationship between the self and the world. Like any communitarian, I am happy to insist that selves gain value only through their encounter with experience and their reflection on their relationships. My disagreement arises only in response to a less metaphysical, but hardly unimportant, question: Who has the right to interpret the authoritative meaning that individuals give to their lived experiences and particular relationships?

For liberals, each person should be recognized as the authoritative author of her own autobiography. The fact that each person draws her life story from a particular community's legacy of language, practice, and tradition does not imply that she is any less the author of her own meanings in life. On the liberal view, the state oversteps its authority when it denies the legitimacy of this crucial act of self-definition.

In contrast, communitarians look with ironic distance upon each person's fragmentary and erratic efforts to construct a life story through ongoing processes of experience, relationship, and autobiographical reflection. Such Sisyphean labors should not blind us to the way individuals may casually destroy the legacy of collective meanings that make their puny efforts possible.

It follows that the communitarian has no principled objection to distributive decisions made in an avowed effort to preserve collective ideals of the good life that sustain the individual's search for meaning. For liberals, the state has overstepped its bounds when it distributes more basic resources to those who are expected to live out their lives in accordance with its scripted versions of meaning in life. Such state-sponsored teleologies offend basic liberal commitments as they are elaborated within the temporal frame each of us adopts at our autobiographical moments.

To put this point in parallel with my others: just as the horizon of lived experience frames a concern with corrective justice which emphasizes the centrality of the *responsibility* of persons, and the horizon of meaningful activity frames a concern with relational justice which emphasizes the *dignity* of persons, I shall say that the autobiographical horizon frames a concern with distributive justice which emphasizes the *freedom* of each person to interpret the meaning of her own destiny.

Even if this liberal framework for distributive justice were accepted, there are the usual difficulties in working it up into an incisive theory. Rawls has one way of proceeding. Nozick has another. I have my own opinions. But at least we are clearer why liberals find the question worth asking.

VI. CONFLICTING TEMPORALITIES

"Responsibility," "dignity," "freedom": I have introduced these three liberal ideals, and the way they emerge from three temporal horizons, one at a time. But when we put the horizons together, do they yield a consistent whole?

This question arises on at least two levels. The first implicates philosophical anthropology: Is there a consistent image of human life that organizes the three aspects of liberal justice as they are refracted by the temporal horizons?

A second involves the nuts-and-bolts construction of one or another theory of justice. As the theorist tries to understand the sources of a particular conflict, she is well advised to place it against the interpretive background provided by each of the three horizons. But each temporal framework may offer conflicting diagnoses of the just resolution of the dispute: ideals of responsibility may point in one direction, principles of dignity in another, notions of freedom in a third. When this happens, how to decide which temporal frame should be most salient?

Taking this second question first, I concentrate on the ways in which *framing paradoxes*, as I shall call them, would arise even in an ideal world—in which all inhabitants were willing and able to build a polity that perfectly complied with liberal principles. In adopting this counterfactual premise, I do not wish to divert attention from the deeply entrenched injustices of the real world. I adopt my idealizing posture only to suggest that even in the unlikely event that liberalism triumphed decisively in the world, it will still have things to worry about.

1. Reframing the meaning of lived experience. Imagine an ideal liberal world in which initial resources were justly distributed and where well-informed entrepreneurs and consumers operated in a competitive market economy.[4] Considering your range of options, you have decided to open a McDonald's and devote your working life to customer service. Picture yourself happily flipping hamburgers when Shmoe enters and deliberately tries to destroy this experience and, with it, the means of your economic livelihood. Stipulate further that Shmoe is just plain nasty: though you have done him no wrong, he has no other mission in life but to make you miserable.

He succeeds: thanks to him, and him alone, your hamburger flipping comes to an end. This is the stuff of corrective justice—however much particular theories disagree on other matters, this looks like a paradigmatic case: Shmoe has done you wrong through intentional actions that have no other motive than sheer maliciousness. Until I add some further facts.

[4] I frame this characterization of the liberal market ideal abstractly to embrace a broad range of particular theories.

Stipulate that Shmoe has not hired a goon squad to scare away your customers, but has opened up his own Hamburger Heaven next door out of pure spite. Driving down his prices, he drives you out of business, chortling loudly as the gas company shuts off your hot plate,[7] and you leave the premises in despair. Do you still think he has done you wrong? Do you think that the liberal state should authorize you to sue him for compensatory (or punitive) damages?

These questions expose a tension between two horizons. From the vantage point of lived experience, the paradigmatic wrong is malicious and intentional disruption. Within the horizon of relationships, the episode has a different meaning—at least if you and Shmoe understand yourselves as engaged in the distinctive activity called entrepreneurship. Within this temporally-extended frame, we expect to be judged over time by how well we take advantage of opportunities opened up by the market—hence it is not wrong if you go under because somebody else has decided to undercut your price.

Even Shmoe's jubilation takes on a different appearance: Just as the baseball players who win the World Series jubilate on the field after their triumph, is there anything really wrong when Shmoe trumpets his own victories in the "game" of free competition? At worst, Shmoe has acted in bad taste, exhibiting emotions hidden beneath a veneer of polite concern by those with better breeding.

Is Shmoe, then, a loud boor or a monster of injustice? Perhaps this is not a problem that cannot be resolved cleanly one way or the other. I shall express this continuing sense of irresolution by saying that thoughtful partisans of one or another answer *suspend* rather than *dismiss* the objections raised by their opponents. Suppose, for example, that liberal citizens resolve the framing paradox by passing a law that allows Shmoe to escape from legal liability: "Whereas people should have the right to relate to one another as competing entrepreneurs in a market economy, they have the right to cut prices even when intentionally inflicting foreseeable harm upon other competitors...."

Most thoughtful advocates of this law will retain a capacity to experience the shock of injustice as they view the malignant Shmoe triumphing over his hapless victim. Thus, they will not find it possible to *dismiss* the proponents of corrective justice who find Shmoe's willful disruption morally wrong. Instead, they want to *suspend* this judgment as they focus on the way price cutting appears within a relational time frame: If entrepreneurs find it meaningful to relate to one another as participants in a market economy, should the liberal state not respect their right to tell this story about themselves? If price cutting is a significant part of this story of competitive prowess, why should the state seek to ban it?

---

[7] Some might be impressed by the fact that it is the gas company, rather than Shmoe, that is the direct cause terminating your job. If you are one of them, stipulate that Shmoe owns the gas company.

On the other hand, those who would pass a statute condemning Shmoe for "cut-throat competition" or "predatory pricing" will also be aware of a distinctive normative undertow. While they may insist that Shmoe is doing something wrong, they may find it difficult to blind themselves entirely to the stories entrepreneurs tell one another as they compete for business in the marketplace. In their view, however, these stories should not allow us to tolerate behavior that would otherwise be condemned within the frame of lived experience.

Although most of us tend to one of these suspensive views, I do not deny the existence of thoughtful people who are downright dismissive. Strong deontologists often suggest that intentional maliciousness can *never* be outbalanced by the beneficial consequences of activities. Unfortunately, they do not test the rigors of their doctrine by looking the Shmoes of capitalism in the face and telling us how they would run an economy that did not tolerate countless acts of intentional, short-run, harm doing. They write as if their critique of utilitarianism only had bite in crisis situations arising in war or other grave emergencies.[8] Perhaps they can maintain the credibility of their rigorous doctrine only by blinding themselves to many of the cases that it would condemn?

This is, at least, the opinion of classical utilitarians, who have recently been reinforced by the American school of lawyer-economists. These analysts turn the tables on deontology by refusing to take seriously the frame of lived experience and by focusing all their attention on the extended frame of meaningful relationship. For them, the legitimacy of the marketplace serves as a paradigm for the analysis of other problems of conflict. If, as they suppose, the short-term cruelties of the marketplace are readily justified by a longer-term view of the benefits, a similarly expansive approach is appropriate when it comes to assessing other areas of social life.

From this perspective, it is naive to condemn a relationship merely because it looks morally unattractive at a particular point in time. Almost every activity worth doing has its unattractive moments. The question is whether these bad patches are outweighed by the good ones that the activity generates at other times. If the answer is 'yes', the right thing to do is to transcend the primitive moralisms of lived experience and embrace the activity with all its imperfections—so long, at least, as you have taken steps to minimize the bad and maximize the good it generates over time. By the simple act of *transvaluing time horizons*, we have shifted decisively away from deontology toward maximizing cost-benefit analysis.

---

[8] See, for example, Judith Jarvis Thomson on trolleys, in *The Realm of Rights* (Cambridge: Harvard, 1990), chap. 7; Bernard Williams on wartime atrocities, in "A Critique of Utilitarianism," in J.J.C. Smart and Williams, *Utilitarianism: For and Against* (New York: Cambridge, 1973), pp. 98-100.

But is it really necessary for us to continue dismissing one another in these familiar ways? On the one hand, the utilitarian is surely right in pointing out the naiveté of fixating on bad moments without giving credit to the good an activity can produce. On the other hand, the deontologist is right to point to the special terror we feel when some rival intentionally disrupts our lived experiences without even attempting to justify his actions by appealing to principles of corrective justice.

Rather than dismiss either of these points, liberal public policy should aim to express the moral insights expressed on both sides of the framing paradox. For example, we should allow entrepreneurs to drive their competitors out of business, but require them to pay a special tax to cushion the unemployment that their intentional behavior predictably causes.

I suspect that similar straddles are also available in other areas of social policy.

2. *Reframing the question of dignity.* A different kind of straddling is in order when confronting a second framing paradox. This one arises when we juxtapose the relational horizon against the life-story perspective we adopt in our autobiographical moments.

I introduce the paradox as it might appear from the vantage of ideal theory. To fix ideas, suppose that each citizen of our ideal state receives a liberal education introducing her to a broad range of life options, that each begins adult life with a roughly equal share of wealth,* and that each is held responsible for disruptive behavior by a system of criminal and civil law rooted in a sound understanding of corrective justice.

Citizen Freedom emerges from her period of tutelage to consider the job choices that are offered up to her in a competitive marketplace. After surveying her full range of choices, Freedom narrows her options down to two: a dignified workplace in which her fellow workers treat her contributions with the respect they deserve and which pays a salary of $35,000 a year; an undignified competitor, in which the little ladies are treated in a polite, but unserious, manner by their male colleagues—even though they are paid $40,000!

---

* I try to justify these, and other basic entitlements, in my *Social Justice in the Liberal State* (New Haven: Yale, 1980). Other liberal theories understand distributive justice differently, but in many cases, these differences will not change the basic structure of the framing paradox of interest here.

This is not as impossible as it may sound. To use economic jargon, all it requires is that some men gain enough "psychic income" from machismo that they are willing to sacrifice enough dollars so as to offer Citizen Freedom a deal on undignified terms.[10] Stipulate further that Freedom acts under conditions of perfect information—she is carefully told what the terms of the deal involve. After weighing the matter, Freedom accepts the offer of undignified employment—taking the bigger paycheck and salving her ego with a fancy mink coat.

Unless, of course, the citizenry intervenes to make this kind of transaction illegal. If the woman had not consented, liberal principles would indeed condemn organizations that stripped her of her dignity. Does the bare fact of her consent make such a big difference?

My own answer emphasizes the distinction between ideal theory and real-world conditions. When considering the question as it would arise in an ideal liberal state, you should not assume that Citizen Freedom has received an illiberal education which sought to indoctrinate her in a life of feminine subordination; nor should you assume that Freedom comes from the lower classes and chooses the higher income because she has been unfairly deprived of a roughly equal starting point in life. These real-world factors do give rise to a very powerful argument in favor of banning dignity-stripping arrangements, even those reached by knowledgeable adults in uncoerced settings. Quite simply, given the existing injustices in the distribution of educational and material resources, do we want to allow Citizen Freedom to reinforce by her conduct the idea that there is nothing wrong with stripping women of their dignity?

But such questions are excluded by definition in my idealizing thought experiment, which challenges us to consider whether, even in an ideal world, Freedom should be barred from stripping herself of her own dignity as a worker by trading it for cash. I answer by affirming the value of freedom over dignity: when push comes to shove, I place the highest value on the freedom of each citizen to organize activities and experiences into a life story that *she* finds meaningful. If Citizen Freedom knowingly sacrifices her dignity on the job to gain resources for the more meaningful pursuit of other activities, it is not for the state to say that she is wrong.

---

[10] This theme is elegantly explored (without too many moralizing words like 'dignity') by Robert Frank in *Choosing the Right Pond* (New York: Oxford, 1985).

Of course, one should make very sure that there is a real conflict between freedom and dignity before allowing the latter to be sacrificed. Consider prostitution—a paradigmatically undignified activity as it is generally practiced. Doubtless, many prostitutes are driven to the activity by unjust conditions of poverty and inadequate education. Others do not know what they are in for. While all this justifies a great deal of regulation, I do not think it justifies an outright ban. If a twenty-year old with a good education decides to make money off her body by selling it to partners (who do not physically abuse her), she should be treated as the ultimate author-ity in the matter. Even in the real world, legitimate concerns for dignity should never eliminate the possibility of freedom.

But let us return to ideal theory, and probe more deeply into the framing paradox. Suppose now that Citizen Freedom goes further, and announces her intention to enslave herself totally and forever to a master. Must the liberal state allow her to strip her claims to dignity so completely?

Many of my friends respond by refusing to answer. After all, they point out, we are dealing in ideal theory, and therefore should suppose that Freedom has begun life with her fair share of basic resources. Under such conditions, it is almost impossible to suppose that she will "voluntarily" choose slavery.

I agree that it is unlikely, but it is always a mistake to underestimate human freedom, and its variety of uses. Ideal theory is ideal in its grant of equal freedom to all; it does not guarantee that citizens make an ideal use of their freedom. Let us stipulate, then, that one of our fellow citizens announces her intentions to become a slave-for-life—if only to expose the shallowness of liberal ideals. Does she have this right to strip away her dignity?

Begin by distinguishing this extreme case from the others we have considered—on two dimensions. First, most dignity-stripping cases allow the stripper to change her mind after a relatively short period of time. Recall my earlier hypothetical case involving a woman worker trading $5,000 in salary for undignified treatment. If she tires of this deal, she remains free to change her life story by changing to a job that promises lower pay for more dignity. Second, most other cases involve piecemeal stripping of dignity in one activity, leaving the citizen free to write a dignified autobiography by emphasizing her other activities and experiences.

In contrast, the slavery-for-life case involves a person who is completely stripping herself of her author-ity. She is denying that she is the sort of creature who takes responsibility for writing an autobiography.[11]

I would draw the line at this point, and refuse to recognize the legitimacy of the action. Of course, if somebody were grimly determined to disclaim all author-ity, there may be precious little we can do in practice to stop her—other than to make it clear that we would prevent the master from maintaining his dominion if she should ever express the desire to take up her life again as a free citizen of the liberal state.

In any event, my aim here has not been to resolve such paradoxes, but to urge you to consider them with greater care.

VII. METAPHYSICAL FOUNDATIONS OF LIBERALISM

In a recent essay,[12] I reaffirmed my faith in political liberalism—so long as one did not take its promises of liberation from metaphysics and epistemology too seriously. While political liberals try, as a matter of principle, to evade contestable matters of first philosophy, they do not suppose they can altogether avoid such confrontations.

This essay has pointed to one of these necessary encounters. No serious theory of justice can evade the ancient dispute between monists and dualists. It arises whenever the debate becomes self-reflexive and considers the basic structure of the idea of justice itself. While monists have the advantage of Ockham's Razor—Why make a difficult subject more complicated than necessary?—dualists succeed in raising nagging doubts about the adequacy of any monistic conception to do justice to our considered convictions.

Rather than repeat pat dualistic formulae, I have tried to reflect upon fundamental aspects of the human condition and show how they may motivate a trinitarian response. Since the problem of justice only arises amongst creatures endowed with time consciousness, I have searched for an answer by elaborating a more general account of the struggle for meaning in time, seeking to show how it might enlighten our particular problem.

I have tried to conduct this search in the spirit of political liberalism—providing just enough metaphysics to motivate a thoughtful answer without indulging elaborate metaphysical conceits that needlessly alienate people who might otherwise embrace liberalism's key commitments. Despite the wild diversity of outlooks held by the citizens of a modern society, are there really lots of people out there who fail to recognize the relevance of the three temporal horizons—lived experience, meaningful relationship, life story—which frame their search for meaning in life? I do not think so.

---

[11] Compare Harry Frankfurt's treatment of the wanton in his "Freedom of the Will and the Concept of a Person," this JOURNAL, LXVIII, 1 (January 14, 1971): 10-14.

[12] "Political Liberalisms," this JOURNAL, XCI, 7 (July 1994): 364-86.

If my thesis is novel, it lies in the effort to link this spare metaphysic to the problem at hand. In making this link, I do not claim that the three temporal horizons are themselves sufficient to solve the trinitarian problem of justice they reveal. Merely because all of us struggle for meaningful experiences/relationships/lives, it does not follow that we all adopt the same principles of justice when our struggles for meaningful experiences/relationships/lives conflict with one another.

As we reflect on the framing paradoxes that ensue, liberals may find it necessary to probe more deeply into the metaphysical foundations of the self's encounter with time. In my brief introduction to framing paradoxes, I have suggested my own strong priority to the self's struggle for a meaningful autobiography over the claims of particular relationships and lived experiences. But I shall leave a more extended defense of this priority to another time.

BRUCE ACKERMAN

Yale University

BOOK REVIEWS

Kant et le pouvoir de juger: Sensibilité et discursivité dans l'Analytique transcendentale de la Critique de la raison pure. BÉATRICE LONGUENESSE. Paris: Presses Universitaires de France, 1993. 482 p.

In the two editions of the *Critique of Pure Reason*, Immanuel Kant claimed that thinking about any object, any possible "representing of an object," was a "judging." Judgings were understood to be rule-governed mental acts, and the formal characteristics of any possible object were said to reflect or depend in some way on the formal rules of judging. His language of mental faculties and mental activities and their "logical rules" has suggested to many generations of readers that, if Kant was trying to justify a priori knowledge about the necessary characteristics of any world we could experience, that argument must have two steps: an appeal (1) to some (implausible) claim about the way the human mind must work, about the necessity of our thinking in certain ways (or some "psychologism"), and (2) to some idealist reduction of "the world" to a world of mental states and mental "products."

This impression is strengthened by Kant's infamous claim for a "strict parallelism" between the rules the mind is subject to in combining, subsuming, or comparing any representations at all (the rules for judging itself, or "general logic") and the rules to which the understanding is subject in making any objective (possibly true) judgment ("transcendental logic"), as if some argument could show that the discursive forms of logical combination as such, when thought together in some way with our modes of sensible contact with the world, determine—could be said to anticipate in some way—what could be a possible content of sensible experience. This again looks like it could only be a "mind making the world" argument.

This claim for a link between general logic and a transcendental logic constitutes Kant's famous "clue" or "guiding thread" (*Leitfaden, le fil conducteur*), the founding claim which begins the "Transcendental Analytic" and which is supposed to help us understand, say, the relevance of the subject-predicate form to the claim that there must be underlying, permanent substance in any possible experience; hypothetical judgments to the necessity that all events are causally connected with others according to a rule; even disjunctive judgments to

0022-362X/97/9406/318–24

reciprocal causation and a community of substances. The guiding thread is summarized this way:

> The same function which gives unity to the various representations in a judgment also gives unity to the mere synthesis of various representations in an intuition; and this unity, in its most general expression, we entitle the pure concept of the understanding. The same understanding, through the same operations by which in concepts, by means of analytic unity, it produced the logical form of a judgment, also introduces a transcendental content into its representations, by means of the synthetic unity of the manifold in intuition in general (A79/B105).

Béatrice Longuenesse's very ambitious new book takes a fresh, comprehensive look at this Kantian theory of judging, and remarkably (given the notorious obscurity of the issues), concentrates very heavily on the above *Leitfaden* problem, the most difficult and obscure implication Kant drew from his theory of judgment. She has thereby produced a rich, often demanding, sometimes elusive and metaphorical, but always provocative and bold approach, one that will likely command and sustain a good deal of attention, especially when the English translation is published next year by Princeton University Press.[1]

Its nearly five hundred pages are also not easy to summarize or to respond to briefly. In order to address the question of *how* Kant could think that consideration of the forms of judgment could "guide" an argument about the necessary conditions of experience, Longuenesse must produce an account of the theory of judgment itself, assess its historical origins and distinctness, anticipate and integrate Kant's later treatment in the third *Critique*, try to make sense of his own architectonic presentation of the table of judgments, and then, especially, try to detect and to follow very closely any possible relevance of claims about judgmental forms to claims about possible objects of experience throughout the "Deduction," "Schematism," and "Principles." I know of no other commentator who has taken Kant so much at his word—that there *is* a guiding thread uniting the *entire* argument of the "Analytic"—or who has produced such a plausible view of what he might have had in mind. Much of what she says in the course of her interpretation is compelling and original. The treatment of the uniqueness of Kant's understanding of general logic (and the differences between his and modern conceptions), and the demonstration of the overall coherence of Kant's account of judgment throughout most of the critical period (especially her demonstration of the implicit presence and power in the first *Critique* of his later distinction between determinative and reflective judgments) are signal accomplishments.

---

[1] The English edition, *Kant and the Capacity to Judge*, will be expanded and revised. In the text above, in order to ensure fidelity to the author's intentions, I have cited the preferred translations from this manuscript (translated by the author, from a draft by Charles Wolfe).

Although details only begin to emerge in the third of the book's three sections, there is also a "guiding thread" throughout Longuenesse's own treatment, a theme that is supposed to pull all the various parts together successfully into one overall argument. That issue is already apparent in the dense passage from Kant cited above: What *could* he have meant by saying that "functions" which give unity to representations in a judgment "also" give unity to the mere synthesis of representations "*in* an intuition"? *How* does whatever "produces" logical form in a judgment "introduce" "transcendental content" into representations? What could such a claim about content even mean?

Longuenesse finally finds what is for her the clue to all such questions in an even more obscure claim hidden in a small section of the second edition deduction, one which Longuenesse unearths and reads backward and forward into a number of other crucial sections, Kant's account of the "*synthesis speciosa*," or the figurative synthesis at B151. This is where, she thinks, the heart of the matter, Kant's overall view of the discursivity-sensibility relation, and so the relation between forms of judgmental discursivity and possible sensible *content*, might be illuminated.

It is certainly true that this approach—trying to follow whatever "guiding thread" an attention to the forms of judgment might actually provide throughout the details of the "Analytic"—illuminates the course and connections in Kant's defense of a priori knowledge in ways quite different from many other commentators. There is usually, at the center of any such commentary, some such orienting concern; a claim that the whole argument finally turns on the possibility of self-ascribing my mental states, or on tying down somehow claims about time determinations, or on being able to distinguish what seems to me from what is the case, and so on. Longueneese's proposed clue (any object of experience, that is, any sensibly given, must be *the kind of content susceptible* to such judgmental organization, and we must be able to explain *how* it could be) provides her with a two-part claim for such a basic orientation. First, since thinking any thought is a judging, we should keep in mind what judging is, primarily what Kant often calls "reflecting" (or "analysis"), an activity most clearly on view in the formation of empirical concepts, but involved in all mental content. Thinking according to Kant is an ordering of concepts "under" each other: a classifying, subsuming, and most especially, Longuenesse shows to great effect, a *comparing* of representations. (In Longuenesse's own view, this stress also sets her interpretation apart because it so emphasizes the "reflective" relation between the sensible and the intellectual, a way of understanding how the sensibly given is "brought up" into an intellectual form, rather than "determined" "from above" by the understanding, as in standard accounts. It thus also leads her to characterize Kant's position on universals in a surprisingly realistic way (see 143).)

Secondly, according to Longuenesse, Kant wants to show that the forms of any such logical relations must play a necessary role in an original "synthesis" of elements suitable for, finally the contents of, such analysis. ("It must therefore be shown that these forms are not only conditions of the subordination of concepts according to logical use, but conditions of the *very presentation of appearances in sensible intuition*, a presentation that generates for these appearances objects reflected under concepts" (16).)

The most important parts of the book are an attempt to explain and defend this last claim. I have space for only a couple of general remarks about the attempt.

Formulations like this one and several others raise questions about the version of *idealism* required by this account, and so the right characterization of the nature of the "objects" of human knowledge. There is no particular reason that Longuenesse's position should necessarily be committed to the claim that the content of knowledge must be ontologically mental—sensory states, say, somehow "preformed" for objective claims by some figurative synthesis so as to be able to function as the objects (= unified sensations) of empirical claims—but her formulations are ambiguous. There is no direct commitment to such phenomenalism, but when she mentions early on that the relation between representation and object had been "internalized" after section 14 of the *Critique*, so that there is "no longer...any direct relation to an object outside representation [extérieur à la représentation]" (9), it is not clear exactly what this leaves us with. One possible account of what it means to say there can be nothing available to us cognitively "outside" human representation is: to represent an object, and so for Kant simply to be conscious of objects, is to be making a judgment with a certain sort of normative force or implicit claim to justification, an *entitlement* that cannot be said to derive from anything that merely "happens" to the subject ("from the outside"), but must derive that force from other sorts of appropriate uses or justifications involved in making such a claim. The object can only be "that in the concept of which the manifold is united" and so the *appropriate* use of the concept cannot be any direct appeal to an (outside) object. But Longuenesse's language here of mental making, generating, forming, imprinting, and "inside" suggests an account of some more literally ontological rather than such a normative "inside" and so a phenomenalist direction and a so-called "two-worlds" idealism. (See the discussion in chapter nine on logic, ontology, and the crucial role of the *synthesis speciosa* again (379).)

The second question concerns Longuenesse's attempt to answer the heart of that "clue" problem: What does it mean to say the functions of unity in the understanding "give unity" "*in* an intuition"? In Kant's second-edition language, this question concerns what it could mean to say that the understanding "determines sensibility inwardly" and this is a much more radical question than the question of how the understanding organizes and connects an already determinate sensory manifold. Some such "inward determination," involved just in the "original" consciousness of a manifold, our way of apprehending it, itself apparently forms a critical component of the deduction's argument that there *could* never be any "lack of fit" between the formal requirements of judgment and the (at least to some extent) independently given manifold. Enter the poetically named *synthesis speciosa*. According to Longuenesse, after the second half of the B deduction, Kant had "radicalized" his deductive procedure far more than has been realized, completely reinterpreting "the manner in which things are given to us" (235), and so had engineered a major reconstruction of the "Transcendental Aesthetic" itself (as Kant seems to imply in several famous footnotes.)

Any such claim about an intellectual "generation" (242) of space and time "and thereby the form of appearances" and so forth must be presented very carefully. It threatens to collapse the understanding-sensibility distinction at the heart of Kant's critique of rationalism (showing that the content of any manifold must be susceptible to our forms of judgment cannot ever "amphibolously" confuse what is required for judgment with what must come to us independent of judgment, else the central pillar, or twin pillars, on which the *Critique* stands would fall), and, as G. W. F. Hegel immediately realized in his early *Glauben und Wissen* (*Faith and Knowledge*), any blurring or relativization of such a distinction immediately opens the door to J. G. Fichte and a very different idealism, wherein the question of what constrains our organization of experience is itself the result of a "positing" or a conceptual determination, one wherein "thought" is ultimately "self-determining."

Longuenesse is well aware of such problems (and of their history both in German Idealism and, especially, in neo-Kantianism) and she works very hard to avoid either a rationalist amphibology or a Fichtean/Hegelian "collapse" (that is, some claim that the original "rendering susceptible" here in question is some sort of conceptual determination). It is not, according to her, because one must read Kant's claims about the role of the understanding in such an original, nonobjective, precategorial synthesis, as only a reference to *the capacity to judge.* It is only this *capacity* to form a judgment that can be said to "affect sensibility," generating the pure intuitions of space and time, the intuitive counterparts to our discursive capacity to form and use universal concepts, as determined by categories (cf. 247-53). This and frequent other references to a subject-constituted "tendency" toward judgment and objectivity in apprehension, or claims that time is the form that inner sense "imprints on what is represented in it" (266), while they might help a reader avoid a number of misleading implications of Kant's statements, are themselves still highly metaphorical. The idea that some "*tendance au jugement*" could "*affecte le sens interne*" (272), that a *tendency* or *Vermögen* could be the sort of thing that could "affect" inner sense, remains, I think, elusive.

This is not to say that a good deal of help is not provided, especially in the book's concluding chapters on the "Principles," which contain some of the most interesting textual analyses, and where the above formulations continue to "guide" especially the account of our capacity both to apprehend and to determine time relations. (So, for example, "according to Kant our own synthesizing acts introduce [*importons*] into intuition the forms of combination we shall eventually reflect in quantitative concepts" (308).) Such claims about the figurative synthesis ground an ambitious argument denying that the general problem of the "Analogies" is, as often assumed, the conditions for objective time determination. Longuenesse argues that that case concerns the possibility of *any* time determination at all and therewith, subsequently, the reflection of perceived orders in universal concepts and laws. Moreover, her approach also makes possible a distinctive way of understanding what Kant could have thought disjunctive judgments could have to do with reciprocal causality and the community of substances in the world, all in a way not illuminated without the *synthesis speciosa.*

We are all familiar by now with "austere" approaches to the *Critique's* basic argument, some reconstruction wherein the issue is only what sorts of distinctions experience must allow us to make in order that some incontrovertible claim or capacity be possible. The motivation for such interpretations has often been suspicion about Kant's mentalist language, and that suspicion is often based on a number of questionable interpretive assumptions. There has been room (and a need) for some time now for a thorough reconsideration of the historical Kant's theory of mental activity, of judging (or his theory of how "*we* make" such distinctions possible), and one which avoids the temptation to read that theory simply in the light of contemporary "cognitive psychology," that is, one which still takes quite seriously his concern with the problem of how we might know a priori that the content of experience must be susceptible, necessarily conforms to, any such mental processing. Such a synthetic claim might not finally be defensible, but such a thorough reconsideration, animated by Kant's greatest philosophical ambitions, and informed by the best erudition and textual fidelity, has now been written.

ROBERT B. PIPPIN

University of Chicago

NEW BOOKS

APPIAH, K. ANTHONY, and AMY GUTMANN: *Color Conscious: The Political Morality of Race.* Princeton: University Press, 1996. 191 p. Cloth $21.95.

ASMA, STEPHEN T.: *Following Form and Function: A Philosophical Archaeology of Life Science.* Evanston, IL: Northwestern, 1996. xiii, 232 p. Cloth $69.95, paper $19.95.

BAARS, BERNARD J.: *In The Theater of Consciousness: The Workspace of the Mind.* New York: Oxford, 1997. xiv, 193 p. Cloth $25.00.

BAKER, GORDON: *Descartes' Dualism.* New York: Routledge, 1996. xiv, 235 p. Cloth $49.95.

BAR-ELLI, GILEAD: *The Sense of Reference: Intentionality in Frege.* Berlin: Walter de Gruyter, 1996. xxv, 251 p. Cloth $124.50.

BARNARD, G. WILLIAM: *Exploring Unseen Worlds: William James and the Philosophy of Mysticism.* Albany: SUNY, 1997. xiv, 422 p. Paper $21.95.

BEABOUT, GREGORY R.: *Freedom and Its Misuses: Kierkegaard on Anxiety and Despair.* Milwaukee: Marquette, 1996. x, 192 p. Paper $20.00.

BEARDSWORTH, RICHARD: *Derrida and the Political.* New York: Routledge, 1996. xvii, 174 p. Paper $16.95.

BEARN, GORDON C.F.: *Waking to Wonder: Wittgenstein's Existential Investigations.* Albany: SUNY, 1997. xxv, 265 p. Paper $24.95.

BEISER, FREDERICK C.: *The Sovereignty of Reason: The Defense of Rationality in the Early English Enlightenment.* Princeton: University Press, 1996. xi, 332 p. Cloth $39.50.

BERGOFFEN, DEBRA B.: *The Philosophy of Simone de Beauvoir: Gendered Phenomenologies, Erotic Generosities.* Albany: SUNY, 1997. x, 250 p. Paper $17.95.

BLUMENTHAL, H.J.: *Aristotle and Neoplatonism in Late Antiquity: Interpretations of the* De Anima. Ithaca: Cornell, 1996. x, 244 p. Cloth $57.50.

BOHMAN, JAMES: *Public Deliberation: Pluralism, Complexity, and Democracy.* Cambridge: MIT, 1996. xi, 303 p. Cloth $30.00.

BOND, E.J.: *Ethics and Human Well-being.* Cambridge: Blackwell, 1996. xiii, 270 p. Cloth $54.95, paper $21.95.

BOYLE, ROBERT: *A Free Enquiry into the Vulgarly Received Notion of Nature.* EDWARD B. DAVIS and MICHAEL HUNTER, eds. New York: Cambridge, 1996 xxxvi, 171 p. Cloth $54.95, paper $18.95.

BRANDT, RICHARD B.: *Facts, Values, and Morality.* New York: Cambridge, 1996. vii, 319 p. Cloth $59.95, paper $18.95.

BUNGE, MARIO: *Finding Philosophy in Social Science.* New Haven: Yale, 1996. xii, 432 p. Cloth $45.00.

BURBIDGE, JOHN W.: *Real Process: How Logic and Chemistry Combine in Hegel's Philosophy of Nature.* Toronto: University Press, 1996. x, 274 p. Cloth $75.00.

BUTLER, JUDITH: *Excitable Speech: A Politics of the Performative.* New York:

Routledge, 1997. x, 185 p. Paper $16.95.

CASEY, EDWARD S.: *The Fate of Place: A Philosophical History.* Berkeley: California UP, 1997. xviii, 488 p. Cloth $45.00.

CLARK, STEPHEN R.L.: *How to Live Forever: Science Fiction and Philosophy.* New York: Routledge, 1995. vii, 223 p. Cloth $74.95.

COOKE, PAUL D.: *Hobbes and Christianity: Reassessing the Bible in Leviathan.* Lanham, MD: Rowman and Littlefield, 1996. Cloth $62.50, paper $23.95.

CRANSTON, MAURICE: *The Solitary Self: Jean-Jacques Rousseau in Exile and Adversity.* Chicago: University Press, 1997. xii, 247 p. Cloth $29.95.

CUDWORTH, RALPH: *A Treatise Concerning Eternal and Immutable Morality.* SARAH HUTTON, ed. New York: Cambridge, 1996. xxxvi, 218 p. Cloth $59.95, paper $19.95.

DANIELS, NORMAN: *Justice and Justification.* New York: Cambridge, 1996. xiii, 365 p. Cloth $59.95, paper $18.95.

DEIGH, JOHN: *The Sources of Moral Agency: Essays in Moral Psychology and Freudian Theory.* New York: Cambridge, 1996. xv, 254 p. Cloth $59.95, paper $18.95.

DELLA ROCCA, MICHAEL: *Representation and the Mind-Body Problem in Spinoza.* New York: Oxford, 1996. xiv, 223 p. Cloth $39.95.

DOUBLE, RICHARD: *Metaphilosophy and Free Will.* New York: Oxford, 1996. xi, 176 p. Cloth $35.00.

EDEN, KATHY: *Hermeneutics and the Rhetorical Tradition: Chapters in the Ancient Legacy and Its Humanist Reception.* New Haven: Yale, 1997. viii, 119 p. Cloth $20.00.

ELGIN, CATHERINE Z.: *Between the Absolute and Arbitrary.* Ithaca: Cornell, 1997. x, 205 p. Cloth $39.95, paper $15.95.

FALES, EVAN M.: *A Defense of the Given.* Lanham, MD: Rowman and Littlefield, 1996. xv, 225 p. Cloth $62.50, paper $23.95.

FLORIDI, LUCIANO: *Scepticism and the Foundation of Epistemology: A Study in the Metalogical Fallacies.* New York: E.J. Brill, 1996. xx, 368 p. Cloth $126.50.

FYNSK, CHRISTOPHER: *Language and Relation.* Stanford: University Press, 1996. xi, 317 p. Cloth $55.00, paper $17.95.

GARRETT, DON: *Cognition and Commitment in Hume's Philosophy.* New York: Oxford, 1997. xiv, 270 p. Cloth 45.00.

GELLMAN, JEROME I.: *Experience of God and the Rationality of Theistic Belief.* Ithaca: Cornell, 1997. x, 211 p. Cloth $32.50.

GENSLER, HARRY J.: *Formal Ethics.* New York: Routledge, 1996. viii, 213 p. Cloth $59.95, paper $16.95.

GILBERT, MARGARET: *Living Together: Rationality, Sociality, and Obligation.* Lanham, MD: Rowman and Littlefield, 1996. xiii, 416 p. Cloth $68.00, paper $22.95.

GOOCH, PAUL W.: *Reflections on Jesus and Socrates: Word and Silence.* New Haven: Yale, 1997. xii, 308 p. Cloth $40.00.

GOWER, BARRY: *Scientific Method: An Historical and Philosophical*

Introduction. New York: Routledge, 1997. vii, 276 p. Cloth $65.00, paper $18.95.

GRAY, JOHN: *Mill on Liberty: A Defence.* New York: Routledge, 1996. xii, 175 p. Paper $17.95.

GRIMES, JOHN: *A Concise Dictionary of Indian Philosophy: Sanskrit Terms Defined in English.* Albany: SUNY, 1996. xiii, 366 p. Paper $26.95.

HAACK, SUSAN: *Deviant Logic, Fuzzy Logic: Beyond the Formalism.* Chicago: University Press, 1997. xxvi, 291 p. Cloth $55.00, paper $18.95.

HAMILTON, RICHARD F.: *The Social Misconstruction of Reality: Validity and Verification in the Scholarly Community.* New Haven: Yale, 1996. xiii, 289 p. Cloth $32.50.

HARE, JOHN E.: *The Moral Gap: Kantian Ethics, Human Limits, and God's Assistance.* New York: Oxford, 1996. x, 292 p. Cloth $45.00.

HARRIS, J.W.: *Property and Justice.* New York: Oxford, 1996. xxvi, 387 p. Cloth $90.00.

HINTIKKA, JAAKKO: *The Principles of Mathematics Revisited.* New York: Cambridge, 1996. xii, 288 p. Cloth $59.95.

HORGAN, TERENCE, and JOHN TIENSON: *Connectionism and the Philosophy of Psychology.* Cambridge: MIT, 1996. xiii, 207 p. Cloth $35.00.

HUGHES, G.E., and M.J. CRESSWELL: *A New Introduction to Modal Logic.* New York: Routledge, 1996. x, 421 p. Cloth $59.95, paper $22.95.

HUTCHINGS, KIMBERLY: *Kant, Critique, and Politics.* New York: Routledge, 1996. xii, 219 p. Cloth $59.95, paper $16.95.

JACQUETTE, DALE: *Meinongian Logic: The Semantics of Existence and Nonexistence.* Berlin: Walter de Gruyter, 1996. xiii, 299 p. Cloth $146.70.

JAMES, PAUL: *Nation Formation: Towards a Theory of Abstract Community.* Thousand Oaks, CA: Sage, 1996. xv, 237 p. Cloth $69.95, paper $22.95.

JAMES, WILLIAM: *A Pluralistic Universe.* Lincoln: Nebraska UP, 1996. xx, 405 p. Paper $15.00.

KAMUF, PEGGY: *The Division of Literature: Or the University in Deconstruction.* Chicago: University Press, 1997. vii, 259 p. Cloth $42.00, paper $16.95.

KANE, ROBERT: *The Significance of Free Will.* New York: Oxford, 1996. viii, 268 p. Cloth $45.00.

KASTELY, JAMES L.: *From Plato to Postmodernism: Rethinking the Rhetorical Tradition.* New Haven: Yale, 1997. viii, 293 p. Cloth $30.00.

KEKES, JOHN: *Against Liberalism.* Ithaca: Cornell, 1997. xi, 244 p. Cloth $29.95.

KOETHE, JOHN: *The Continuity of Wittgenstein's Thought.* Ithaca: Cornell, 1996. xv, 188 p. Cloth $29.95.

KORSGAARD, CHRISTINE M.: *Creating the Kingdom of Ends.* New York: Cambridge, 1996. xviii, 442 p. Cloth $64.95, paper $22.95.

KRISHNA, DAYA: *The Problematic and Conceptual Structure of Classical Indian Thought about Man, Society and Polity.* Delhi: Oxford, 1996. xi, 205 p. Cloth $21.95.

KRISTJANSSON, KRISTJAN: *Social Freedom: The Responsibility View.* New York: Cambridge, 1996. xi, 221 p. Cloth $49.95.

LANDESMAN, CHARLES: *An Introduction to Epistemology.* Cambridge: Blackwell, 1996. viii, 215 p. Paper $21.95.

LANG, BEREL: *Heidegger's Silence.* Ithaca: Cornell, 1996. xi, 129 p. Cloth $19.95.

LE POIDEVIN, ROBIN: *Arguing for Atheism: An Introduction to the Philosophy of Religion.* New York: Routledge, 1996. xxiii, 159 p. Cloth $55.00, paper $15.95.

LEIBNIZ: *New Essays on Human Understanding.* PETER REMNANT and JONATHAN BENNETT, eds. New York: Cambridge, 1996. cxix, 528 p. Cloth $59.95, paper $21.95.

LEVY, DONALD: *Freud among the Philosophers: The Psychoanalytic Unconscious and Its Philosophical Critics.* New Haven: Yale, 1996. ix, 189 p. Cloth $25.00.

LLOYD, G.E.R.: *Aristotelian Explorations.* New York: Cambridge, 1996. ix, 242 p. Cloth $49.95.

LONG, A.A.: *Stoic Studies.* New York: Cambridge, 1996. xvi, 309 p. Cloth $59.95.

LOEWY, ERICH H.: *Moral Strangers, Moral Acquaintance, and Moral Friends.* Albany: SUNY, 1996. xii, 251 p.

LYCAN, WILLIAM G.: *Consciousness and Experience.* Cambridge: MIT, 1996. xviii, 211 p. Cloth $35.00.

LYNCH, AARON: *Thought Contagion: How Belief Spreads through Society.* New York: Basic, 1996. xi, 192 p. Cloth $24.00.

MARTINICH, A.P.: *Philosophical Writing.* Cambridge: Blackwell, 1996. vi, 200 p. Paper $16.95; and *Thomas Hobbes.* New York: St. Martin's, 1997. ix, 156 p. Cloth $39.95, paper $18.95.

MATTHEWS, GARETH B.: *The Philosophy of Childhood.* Cambridge: Harvard, 1996. viii, 136 p. Paper $12.95.

MAUND, BARRY: *Colours: Their Nature and Representation.* New York: Cambridge, 1996. xvi, 247 p. Cloth $49.95.

MAY, LARRY: *The Socially Responsive Self: Social Theory and Professional Ethics.* Chicago: University Press, 1996. xii, 209 p. Cloth $34.95; paper $15.95.

MCDONAGH, EILEEN L.: *Breaking the Abortion Deadlock: From Choice to Consent.* New York: Oxford, 1996. xiii, 280 p. Cloth $49.95, paper $24.95.

MCPHERRAN, MARK L.: *The Religion of Socrates.* University Park: Penn State UP, 1996. xii, 353 p. Cloth $35.00.

MESSERLY, JOHN G.: *Piaget's Conception of Evolution: Beyond Darwin and Lamarck.* Lanham, MD: Rowman and Littlefield, 1996. xiv, 166 p. Cloth $52.50, paper $21.95.

MIDGLEY, MARY: *The Ethical Primate: Humans, Freedom and Morality.* New York: Routledge, 1996. ix, 193 p. Paper $14.95.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIV, NO. 7, JULY 1997

CALCU

THE MATHEMATICAL STRUCTURE OF THE WORLD: THE WORLD AS GRAPH[*]

Many theories of logic and metaphysics since Plato and Aristotle (in fact, since Heraclitus) have hinted at claims concerning relations as well as the "connectedness" of all things. Nevertheless, few theories have forthrightly, and with rigor and clarity, addressed the jointly metaphysical and logical issues of relations and this supposed connectedness—in large part because developed theories of relations were not available until the nineteenth century, after work by Augustus De Morgan, C. S. Peirce, Ernst Schröder, and less self-consciously, Gottlob Frege. Although many philosophical theories have mixed logical observations with metaphysical ones, especially in this century, few if any metaphysical theories have made their treatment of relations a touchstone.

I. THE SHORTCOMINGS OF LOGIC AND "LOGICAL" METAPHYSICS

The view of metaphysics I propose is relational and holistic: the concrete world is a single, large structure induced by a single, two-place, symmetric relation, and thus best analyzed as a certain sort of graph. Every concrete entity "in" the world is a part of this structure and is a structure (subgraph) in its own right. Such entities are individuated (and hence contemplated) solely by their graph-theoretic structural features. A motivation for this admittedly strange proposal begins, first, by claiming that both reality and thoughts are *structures* of cer-

---

[*] An earlier version of this paper was presented at the September 1994 Marvin Farber Conference on "The Metaphysics and Epistemology of Relations" at SUNY/Buffalo, organized by Kenneth Barber, at the SIGMA XI Society at SUNY/Fredonia, and in 1997 to the Department of Logic, Charles University (Prague). I thank Michael Grady, H. Joseph Straight, Morton Schagrin, Kenneth Lucey, W.J. Rapaport, Peter Hare, Fabio Morales, Barry Smith, Don Roberts, D.M. Armstrong, James A. Davis, and Darry Johnson for comments. I especially thank Wallace Matson and Tibor Machan for comments and encouragement.

0022-362X/97/9407/329–58

tain sorts and, then, by arguing that the correct or most perspicuous portrayal of this structure is purely relational and, in fact, best portrayed by *graphs*.

There are components of this view in the relationalism of Peirce, in the metaphysical holism of Baruch Spinoza, and in what we could call the "structuralism" of metaphysically inclined set theorists and mereologists. Karl Popper advocated at least a methodological relationalism, and David Mertz has taken logico-metaphysical relationalism seriously.[1] T. L. S. Sprigge[2] has recently proposed a doctrine of "holistic relations" in his examination of F. H. Bradley's theory of relations. Among recent authors, D. M. Armstrong[3] has most clearly considered some of the issues I shall investigate. His conclusion is that every fundamental property could ultimately be a "relationally structural property" (*ibid.*, p. 8). A whole school of researchers in artificial intelligence and cognitive science have developed positions in "knowledge representation" that utilize diagrams and networks of conceptual relations. The proposal with which I am most familiar—and which partly inspired my own approach—is the semantic network processing system (SNePS).[4]

Finally, a number of researchers have suggested that traditional logical representations are inadequate, or misleading, in portraying conceptual and especially "perceptual" information. Jon Barwise and others[5] have suggested this view and urged the importance of diagrams in teaching, learning, and thinking about logical structures. Connectionist theories of mind, and neurological or neural-network analyses of mental processes and states, have still more strongly urged a "network" approach to these phenomena.

My own proposal differs from these diverse suggestions in at least three ways. First, mine is a *metaphysical* proposal: it is a theory about the world and its structure, not just about how the mind or brain is organized. Second, I do not take diagrams and two-dimensional structures as merely notationally or pedagogically convenient tools for portraying traditionally "logical" structures: I propose diagrams, graphical structures, as a serious alternative to logic as it has been traditionally conceived. Finally, I seek to return metaphysical structure to something that can be clearly and indeed mathematically thought about and discussed, and thereby avoid both a merely suggestive holism and the pseudo-mathematical formalisms of logic or networks, while retaining their respective assets.

---

[1] Popper, *Popper Selections* (Princeton: University Press, 1983), pp. 165–66; Mertz, *Moderate Realism and Its Logic* (New Haven: Yale, 1996).

[2] *James and Bradley* (La Salle, IL: Open Court, 1994).

[3] *A Theory of Universals* (New York: Cambridge, 1978).

[4] John Sowa, ed., *Principles of Semantic Networks* (San Mateo, CA: Morgan Kaufmann, 1991), *Conceptual Graphs for Knowledge Representation: First International Conference on Conceptual Structures, ICCS'93* (New York: Springer, 1993); S.C. Shapiro and W.J. Rapaport, "SNePS Considered as a Fully Intensional Propositional Semantic Network," in N. Cercone and G. McCalla, eds., *The Knowledge Frontier: Essays in the Representation of Knowledge* (New York: Spring, 1987), pp. 262-315; and W.J. Rapaport, "Logical Foundations for Belief Representation," *Cognitive Science* x (1986): 371-422.

[5] Barwise and Gerard Allwein, eds., *Logical Reasoning with Diagrams* (New York: Oxford, 1996); J. Glasgow, N.H. Narayanan, and B. Chandrasekaran, *Diagrammatic Reasoning: Cognitive and Computational Perspectives* (Cambridge: MIT, 1995).

We may, crudely to be sure, divide an ontology into two parts: the kind of entities proposed to exist, an *ontological inventory* if you will, and a *structural system* or theory of structure by which these entities are constituted and related. In most materialist systems, for example, the ontological inventory consists of "material objects," while the structural system consists in their organization by spatiotemporal relations.

Much ingenuity has gone into devising subtle, alternative proposals for materialist and nonmaterialist ontological inventories. Alas, remarkably little energy has gone into developing articulated, alternative systems of ontological structure. Other than materialist spatiotemporal models, the primary program for ontological structure in the twentieth-century analytical tradition centers on logical structure. This latter approach is the strongest current in analytical metaphysics, is epitomized by the extreme position of logical atomism, and is crystallized in Rudolf Carnap's famous title, *Der logische Aufbau der Welt* (1928) (*The Logical Structure of the World*). By *logics* of the sort on which philosophical analysis has been based, I mean the following. First, they are formal, and usually symbolic theories, built from symbolic (categorematic) constituents intended to have metaphysical correlates, combinations of which form the most basic, complete ingredients necessary for the description of the concrete world: sentences, propositions, states of affairs, or situations.[4] Second, they have uniformly presumed that there exist basic entities—individuals—that "have" properties, which are reflected in sharply varying accounts of individual constants (proper names, rigid designators, *Urelementen*, and so on) and of the predicates that apply to them.

---

[4] Another feature of most logics is that they are "discontinuous": the grammar of syntactic items and metaphysical status of their correlates involves a break between terms, on the one hand, and formulae made up of terms, on the other, or between individuals and predicates, and sentences formed from them. The view I am proposing is, perhaps like the Lockean use of "idea," continuous: there is no metaphysical demarcation between the nature (or concept) of a thing and of a state of affairs or sentence.

Third, "logical" structure is portrayed by strictly linear strings of symbols that both capture the content of, and have the same superficial form as, natural language. Second-order, "deviant," modal, and intensional logics, and even set and mereological theories can all be seen as variations of this basic approach. With respect to the crucial second feature, involving a contrast between individuals and properties, even Aristotle's metaphysics counts as a logical theory.

There is a widespread presumption that these logical theories are now perfectly well understood in the sense that a mathematical theory is, such as number theory, Euclidean geometry, or the theory of groups. The "term" logics of Aristotle, G. W. Leibniz, and G. Boole articulate the straightforward structure of a Boolean algebra, but the more recent quantified predicate logics—and especially set theory—have, by comparison with theories originating in pure mathematics, such as group theory, a very messy and ill-understood mathematical structure. It is thus perhaps understandable that many mathematicians have been suspicious of a "foundational" approach to their discipline that roots crisp, well-understood mathematical structures in poorly understood theories with comparatively large numbers of complexly interacting axioms, such as set theory. I would maintain that the very possibility of a clear understanding of the world requires the possibility that it *is* a simple mathematical structure, and that creating complex, ad hoc, or hybrid structures for this task constitutes negative progress.

Logics, as well as their extensions in set and mereological theories, have been especially dominant in the twentieth century as proposals for metaphysical and conceptual structure. The various alternative set theories seem to have enormous expressive capacity, and have attracted serious proponents of such views as that there *exist* sets (of simple things, often material ones), or even (W. V. Quine) that there exist *only* sets. This latter view has the attractive feature that every existent's only features are simultaneously its structural features as described by "the" axioms of set theory. Unfortunately, set theory is, in terms of its own mathematical structure, a notorious and ill-understood behemoth, about which there is much quibbling, artfully hidden from and by many philosophers, over "the" nature of sets as described by various proposed axioms of choice, infinity, foundation, the continuum hypothesis, and so on.[7]

---

[7] I allude here to critical histories of set theory, such as M. Hallett, *Cantorian Sets and Limitation of Size* (New York: Oxford, 1986); S. Lavine, *Understanding the Infinite* (Cambridge: Harvard, 1994); and my "Peirce's Philosophical Conception of Sets," in N. Hauser and J. van Evra, eds., *Studies in the Logic of C.S. Peirce* (Bloomington: Indiana UP, 1997).

One should, of course, not underestimate the enormous progress that has been made in the last two centuries within this logico-metaphysical tradition. One goal of set theory has been to "unify" the language and deductive techniques of the sciences, or at least mathematics.[8] Because of its now advanced technical sophistication, the sheer investment put into it by logicians and philosophers to date, a certain pragmatic measure (loosely speaking) of "success" in dealing with problems, and a certain accrued historical legitimacy—none of which a "start up" proposal like mine is likely initially to have—philosophers have not exactly driven themselves to attack logic's premiere status as a structural system. There is also a certain sense—which I shall not contest—that, if any articulated, formal theory of a metaphysical structural system is possible, then a logic is in certain ways "adequate" to express it.

We can, however, observe the following weaknesses in logic as a structural system for metaphysics (and cognition).[9] First, much of our thought seems to be directed toward, and to manipulate, aural and visual structures, or similarly textured entities. These do not seem to be *in* a logical form, even if they can be wrestled into it. A logical coding of this content—such as digitalization—may indeed be possible, but is often unnatural and unperspicuous in exhibiting structure. Likewise, we can express an English sentence as a numerical structure, such as through the trick of identifying positive whole numbers with sentential structure (for example, Gödel numbers or Boolos's simpler scheme), but this does not entail that arithmetical structures display clearly semantic, syntactic, or pragmatic linguistic structure. This objection focuses on logic's portrayal of logical form by linear sequences (strings) of predicates, individual constants, and logical symbols. Such an observation seems to be at the heart of gestures toward two-dimensional logical diagrams that we see, for example, in Peirce and even Frege,[10] and more recently in semantic-network representations of logical structure. The world might have structure—and surely does have structure if it is graspable—and yet this might not be a logical structure: it has patterns that are "mathematical" but not perspicuously rendered into traditional logical notation. Perhaps the most important such features of experience—for example, music or visual beauty—have a definite and easily cognized structure which is not helpfully described as "logical" structure—or in which their value is not preserved in logical transcription.

---

[8] This unification of science and mathematics seems to be the chief desideratum of Quine's program. How much set theory has actually helped us to understand branches of science and mathematics as "one" is questionable, but there is little doubt that set theory has made possible the uniform description of various mathematical structures. I describe below one of the undesirable consequences of such descriptions.

[9] Compare Mark Wilson, "Can We Trust Logical Form?" this JOURNAL, XCI, 10 (October 1994): 519-44.

[10] For Peirce, especially in his Beta and Gamma existential graphs, this diagramatic structure is genuinely metaphysical (or at least conceptual in being "semiotic") structure. Other logical diagrams, such as those of Venn, Euler, Lambert, Leibniz, and Sturm, seem to be merely pedagogical. See my "History of Modern Logic," in the article "Logic" in the Macropedia of *Encyclopedia Britannica* (1994), and "Logic Machines and Diagrams," in the *Routledge Encyclopedia of Philosophy* (forthcoming).

Second, logical structure is historically associated with highly conceptualized thought, and in fact with thoughts that are easily expressed in natural languages. The theory of logical form has thus imitated the structure of this linguistic surface. Such a methodology begins with Aristotle, in which his conception both of logic and of metaphysics is intimately connected to Greek grammar. Even the rare concern that has been expressed about this feature, such as Bertrand Russell's worries about Western subject-predicate forms, or efforts to develop an adverbial account, often merely appeal to (other) features of (other) natural languages. Such linguistic-representational interests may have twisted logic away from its core concern—which I, with Aristotle, Frege, and Peirce at their best, take to be the analysis of demonstration and, more narrowly, deduction. Following other such interests may ultimately make logic a good theory of the linguistically expressible, a good philosophy of language, but a poor theory of the metaphysical structure of the world, or of nonlinguistic thought and mental content. It would be very surprising indeed if one single formal theory, such as first-order predicate logic, would simultaneously be the ideal theory for the theory of demonstration, a foundation for all of mathematics, ideal for a theory of both linguistic and all mental content, *and* for an account of metaphysical structure of the world.

Third, current logical notations, as linear sequences of symbols or "strings," seem to be irredeemably awkward, or even inadequate, at representing certain quantificational phenomena. One such feature is indicated by the phenomenon of branched quantifiers.[11] Quine[12] has remarked upon the awkward nature of all commonly-used notations for quantifiers and other variable-binding operators. Such inelegances, while initially appearing merely to be aesthetic shortcomings in isolated areas, may indicate deeper difficulties.[13]

---

[11] Cf. Quine, *Philosophy of Logic* (Cambridge: Harvard, 1986, 2nd ed.); and Jaakko Hintikka, *Philosophy of Mathematics* (New York: Cambridge, 1996) for a system Hintikka calls independence friendly ("IF") logic that allows the specification of nonlinear dependence of quantifiers and even connectives.

[12] "Peirce's Logic," in K.L. Ketner, ed., *Peirce and Contemporary Thought* (New York: Fordham, 1995).

Fourth, the notion of a logical individual, the notion of a distinct and identifiable thing that "has" properties or enters into relations in either its metaphysical or cognitive aspect—that is, what it is that individual constants or proper names denote—has been enormously problematic. Logical theories, and especially our now dogmatic introductions to logic, often casually equate them with middle-sized, common-sense objects, or even with persons. We might read that New York City or George Washington are "individuals" for logical purposes. Yet these "ordinary" entities—grotesque and naive reifications of a supposedly common-sense metaphysics—are widely known to have enormous internal structure, and are most likely to be best analyzed as composites made up of other "individuals" in more sophisticated thinking.[14] Taking logical individuals to be the smallest subatomic particles that we have identified is only slightly less naive. The most serious, open, and sustained debate over this issue took place within the logical atomism of Russell and the early Ludwig Wittgenstein. But their efforts were eventually regarded, even by themselves, as having been grossly inadequate: Russell was left conjecturing that there existed logically perfect individual names, even if he could not otherwise describe them, while Wittgenstein came to regard this difficulty as crucial and insurmountable, and thus to cast the whole approach aside.[15] This difficulty with the nature of individuals has come to be regarded strictly as a problem for logical atomism. In more recent logico-philosophical accounts, this fundamental problem has simply been ignored, not solved. I wish to suggest, however, that it is a very basic and serious problem with (almost any) logic as a philosophical scaffolding for a theory of world or thought: What exactly are the individuals that have properties? What *are* they and how are they individuated? The ultimate members of sets, not themselves composite sets (if there are any), pose a similar and largely ignored puzzle.

Fifth, logic provides only the framework of a general series of proposals for the structure of thought or reality. Predicate logic does not answer, or even frame the question, for example, of *which* one-place properties are basic, and which reducible, or which (if any) two-place relations are basic, and so on. Furthermore, modern predicate logic has been remarkably blasé about questions of reducibility among the arities of various properties: Are basic one-place properties sufficient (exclusive monadism)? Are basic two-place relations required or perhaps even sufficient (exclusive dyadic relationalism)? Are basic three-place relations reducible to two-place ones? And so on. We might call this feature of predicate logic *polyadism*. Set theory and mereology offer some guidance in the polyadic wilderness and hint that a two-place, nonsymmetric relation is required, and perhaps even sufficient. In set theory, nonmembership relations of various arities, including beyond two-places ones, can be expressed—or at least extensionally described—by means of various Wiener-Kurotowski devices.[16] Peirce proposed that three-place relations are required, but that higher-order relations are reducible to relations of less than four places;[17] he seems to have had no concrete proposals for which one-, two- and three-place relations are basic. In short, modern logic has encouraged, through its arity-neutrality, logic-driven metaphysics to avoid facing certain basic structural questions about properties and relations.[18]

---

[13] For a discussion of how such ambivalences in choosing logical reductions may be symptoms of deeper problems in a logical theory—and specifically with regard to the status of relations in set theory—see my "Set-Theoretical Representations and Their Adequacy for the Logic of Relations," *Canadian Journal of Philosophy*, XI (1982): 353-74.

[14] A similar view is voiced in Sprigge.

[15] See R. Monk, *Ludwig Wittgenstein: The Duty of Genius* (New York: Penguin, 1990). The exact nature of individuals was viewed as an acute problem for late Boolean logic in the works of Peirce, Schröder, and O.H. Mitchell.

Finally, logic gives nonperspicuous accounts of large and important structural features in the world—the organization of the planetary system or of a Ludwig Beethoven symphony. Most such large phenomena would be treated in logic as giant conjunctions (of "facts" about "individuals"), while reality and cognition of them seem to involve features that incorporate holistic and hierarchical notions of pure relational structure which do not always attend to the exact nature of the ultimate constituents (for example, particles in the planets or the pitches in a symphony). Our thinking should be more, and more rigorously, directed toward the larger and more significant structures of the world, not at the bolts and welds of isolated connections, hoping that depth and insight will somehow emerge if we inspect enough of them.

---

[16] See my "Set-Theoretical Representations and Their Adequacy for the Logic of Relations."

[17] Robert W. Burch, *Peirce's Reduction Thesis: The Foundations of Topological Logic* (Lubbock: Texas Tech UP, 1991). Further literature and other issues are discussed in an unpublished manuscript, "Reduction to a Dyadic Predicate Revisited" by A.P. Hazen, provided to me by the author.

[18] See, for example, the polyadic metaphysics in "The Nature of Relations," in R. Chisholm, *A Realistic Theory of Categories: An Essay on Ontology* (New York: Cambridge, 1996), pp. 51-54.

II. CONSIDERATIONS IN FAVOR OF RELATIONALISM

If exclusive monadism is the position that reality requires for a description of its structure only one-place basic properties, I shall call *relationalism* the view that it requires at least one two- or higher place basic relation. Exclusive relationalism would be the view that two- or higher place relations are basic and thus required for a structural account of the world, and that no one-place properties are required. Exclusive monadism is a common view in the history of philosophy, endorsed by Plato, Leibniz, and many others.[19]

In this section, I shall present a number of considerations that suggest that relationalism is true and that relations are necessary for a structural account of the world. It also contains suggestions that monadic properties are derivative phenomena, that is, arguments for exclusive relationalism.

Numbers. It is now quite common to think of the natural numbers, together with "logical" tools for describing patterns among them (for example, functions or sets[20]), as the basis of all mathematical, or at least all quantitative, thought. It is also natural to think of the natural numbers as "entities" in their own right—witness our names for them and the appeals of naive set theory and even of mathematical (better: arithmetical) Platonism.

With the advent of the Peano postulates, and even more clearly in the formalization by Kurt Gödel, we see that a single given number, even the ostensibly unique zero, quite remarkably lacks, within the most sophisticated of our formal theories, identifying (monadic, one-place) *properties.* A number's nature is instead exclusively specified by relationships with other similar entities, notably by the successor relation. In a sense, we cannot say what numbers *are*—at least not without reference to other numbers. This difficulty can be exposed most notoriously through the fact that 3, 5, 7... equally well constitute a "model" of the entities described in the postulates as do the intended 0, 1, 2, 3,...; namely, in the former model, 3 plays the role of the entity that is no number's successor within this set, a structural "zero," and "next odd number" plays the role of the successor.

---

[19] For Plato's views, see H.-N. Castañeda, "Plato's *Phaedo* Theory of Relations," *Journal of Philosophical Logic,* I (1972): 467-80; and a contrary view in David Gallop, "Relations in the *Phaedo,*" *Canadian Journal of Philosophy* Supplement, II (1976): 149-63. For Leibniz's views, see Castañeda, "Leibniz's 1686 Views on Individual Substances, Existence, and Relations," this JOURNAL, LXXII, 19 (November 6, 1975): 687-90; J.A. Cover, "Relations and Reduction in Leibniz," *Pacific Philosophical Quarterly,* LXX (1989): 185-211. For a glimpse at medieval views, see M.G. Henninger, *Relations: Medieval Theories 1250-1325* (New York: Oxford, 1989). Bertrand Russell's apparent relationalism ("we cannot prove that there are such entities as qualities,...whereas we can prove that there must be relations...") is discussed in *The Problems of Philosophy* (New York: Oxford, 1912), p. 95, and elsewhere. On reductions of relations, see Castañeda, "Relations and the Identity of Propositions," *Philosophical Studies,* XXVIII (October 1975): 237-44; I.L. Humberstone, "Comparatives and the Reducibility of Relations," *Pacific Philosophical Quarterly,* LXXVI (June 1995): 117-41; and Milton Fisk, *Nature and Necessity* (Bloomington: Indiana UP, 1974).

[20] The philosophical implications of proposals in category theory are less well established. See Colin McLarty, "Axiomatizing a Category of Categories," *Journal of Symbolic Logic,* LVI, 4 (December 1991): 1243-60, and "Category Theory in Real Time," *Philosophia Mathematica,* II, 1 (1994): 36-44.

This cluster of observations leads us to the hypothesis that all formalizable accounts of "the" numbers, such as the Peano postulates, identify only the relations among numbers—the *numerical structure*, so to speak—not unique properties "true of" numbers but not of other things. There are no distinctive "things" that are numbers, anymore than there is a distinct class of entities that are, by their very nature, "larger than" other things. There are *numerical relationships*; there are various collections of entities—many of them—that exhibit this structure. The system of things with these relationships is itself a structure (a structure-abstract), and *this* is what the number system *is*. We might even reify the structure-abstracts themselves, but must keep in mind that they have no distinctive monadic properties other than derivative ones in virtue of their structural nature.

To put the matter another way, numbers are (just) nodes in certain kinds of patterns.[11] They are like the internally indistinguishable *Urelementen* of Zermelo-Frankel set theory. "Numbers" are properly identified not by what they are but by their relationships to other things. Their identity consists entirely in their "external" relationships. Our reification of numbers—talking of their "existence," "properties," and so on—is a convenient shorthand, of the "place" they occupy in a system of prototypical "things" ordered in certain ways.

Spatial location and other "physical" properties. Leibniz's relational view of space has had a special place in modern science. It has motivated whole research programs in physics, including Ernst Mach's and Albert Einstein's. It is the view that the "location" of things—and their velocities, accelerations, and so on—are not "absolute": they are not objective, intrinsic properties of things.[22] Instead, locations, sizes, velocities, accelerations, and all other spatiotemporal qualities of things are properly to be understood as shorthand for much longer relational descriptions: 'North America' means to the "north" of South America, "west" of Europe, and the locations of these entities are in turn relationally interdefined. Our usual nomenclature and "common-sense" view of such things retain an absolutist view that spatiotemporal properties are one-place, objective, and intrinsic. No one doubts that it is often quite useful to continue to think and speak in this manner, although many theorists now claim that the "real" nature of things underlying our quaint speech and thought about space and time has a quite different structure: it is *purely* relational.[23] Now, since virtually all of our common-sense thought about the physical world involves implicit or explicit reference to spatiotemporal properties, and since our best available scientific accounts rely heavily upon them, it follows that a large part of common-sense and scientific pictures of the world is purely relational (if space-time is): things are "where" they are in virtue of their (spatial) relations to other things.

---

[11] Cf. M. Resnik, "Mathematics as a Science of Patterns: Ontology and Reference," *Noûs*, xv (1981): 529-50; "Mathematics as a Science of Patterns: Epistemology," *Noûs*, xvi (1982): 95-105; as well as remarks in his *Frege and the Philosophy of Mathematics* (Ithaca: Cornell, 1980). See also Charles Parsons, "Structuralism and the Concept of Set," in W. Sinnott-Armstrong, ed., *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus* (New York: Cambridge, 1995). Resnik talks suggestively of a similar general philosophical account of numbers, and informally uses notions of "structures" and "patterns." He seems to lack a precise definition of these notions which I take myself to be supplying in the account of graphs and structure below.

We might still believe that *some* aspects of concrete, experienced things, and our thoughts about these things, are the way they are due to one-place properties, and not just to relational properties. Pondering the common sensibles of Aristotelian metaphysics (the primary qualities of John Locke) at a common-sense level, we might think of the "shape" of a thing as intrinsic and one-place. But shape is, again, a set of spatiotemporal relations, albeit "internal" to the object. Still worse, even to mark out an everyday thing—an apple, for example—from the rest of the world, from its "background," is something we usually do effortlessly and unconsciously through perception. But to perceive an object, to consider a thing *as* an object, is to notice or attribute *contrasts* of various perceivable "properties" against a background. An apple is considered by us an "object" because we can see its outline, and because it is easy to "separate" it from its surroundings by moving it. These contrasts and separability that appear so phenomenologically immediate and monadic (and are the bane of computer optical recognition) are however *really* relational phenomena: the apple is red *compared with the brown ground on which it lies*,[24] and it is easily and separately movable *from the grass and ground*. Thus, even when an aspect of an object does not seem overtly spatiotemporal, and seems monadic (being a distinct object, having a shape), our perception of material objects seems to depend crucially on—or even be entirely constituted of—relational features. Both its monadic properties, and its very objecthood, seem rooted in relations.

---

[22] L. Sklar distinguishes location, velocity, and acceleration absolutism, rejecting location absolutism but proposing velocity or acceleration absolutism—*Space, Time, and Space-Time* (Berkeley: California UP, 1976).

[23] For a survey of the literature on Leibniz's views and relationalist accounts of space see J. Earman, *World Enough and Space-Time* (Cambridge: MIT, 1989) and Andrew Newman, "A Metaphysical Introduction to a Relational Theory of Space," *Philosophical Quarterly*, xxxix (1989): 200-20.

At quite a different level and kind of understanding of material objects, our best scientific theories attribute various "properties" of electrical charge, mass, spin, and so on to the finest microspatial (microphysical) constituents of reality. But how we come to attribute these "properties" to these "things" is only through what we might call their *interactivity*. The notion of a "particle" in physics has difficulties precisely analogous to those of an "individual" in logic. Traditional explanatory systems have an understandable penchant for such footholds in analysis and explanation. It is not difficult to see, however, that our regarding a particle to *have* the "property" of a certain mass is our explanation of why it interacts in certain ways with other similarly interactive entities. We should perhaps instead express ourselves in terms of the root phenomenon, rather than its convenient monadistic shorthand, and say that certain entities interact with other entities in certain ways: this relational interactivity (and a "disposition" for this interactivity over "time") *is* the underlying phenomenon. Objects supposedly having masses, charges, spins, and so on are much like objects having "locations": they are our ways of handily referring to deeply relational phenomena using conveniently monadic expressions.

Words, languages, and meaning. It is a long-standing custom to speak of a word or phrase as *having* meaning; this everyday approach, of treating words as having semantic and syntactic (monadic) properties, has been aided and abetted by theories of language since the Middle Ages. As is often briefly noted but rarely developed, however, this is quite misleading: no series of marks, or of sounds, really—intrinsically—has a meaning. A word *has* a meaning only in the bosom of a language. And the concept of a natural language itself refers to a vast and conceptually ill-understood "system" of teachings, understandings, practices, and even past or present explicit agreements, beliefs, and intentions across a certain social group. In practice, the "meaning" of a word also typically involves its relationship to other words, with their meanings. (This position is broader than what Michael Devitt[illegible] calls "semantic holism.") It is difficult to imagine that a word such as 'gift' has (roughly) the meaning in some language that it does for us without there being a word or expression in that language with the meaning 'possession' has for us; similarly for 'swim' and 'water' (or *some* fluid). Furthermore—and this is still less dwelt upon—what counts as a "word" or morpheme (spoken, written, or other) is itself a highly, if not exclusively, relational notion: it is to have a visual or audible contrast with a "background," to be identifiable as a certain "type"—itself a very difficult notion rooted in "similarity" to other tokens and differentiation from (tokens of) other types. This loose talk about "words" somehow involves shapes of physical object or images, or involves pauses between sounds, thus utilizing a physical notion whose relationality we have already discussed, or in phonological contrasts and relations of pitch, voicing, and so on. Here ("contrastive" phenomena in linguistics), as in music, we note the importance of sound *relations* more than "properties" of sounds. In music, it is pitch relations, not pitch properties, that determine the salient features of identity and similarity of theme, melody, and harmony in pieces, heard in a context of a scale and style system. Our conclusion may well then be that the features of words (indeed, the very identity and individuation of "words") are extensively, even thoroughly, relational: a word exists (has identifiable character as a distinct word), and means, signifies...what it does because of its relation to other words, activities, and things.[illegible]

---

[24] See also E.W. Averill, "The Relational Nature of Color," *Philosophical Review*, CI (1992): 551-88; as well as Russell, pp. 95-102.

Summary: considerations. In our discussion of these three classes of disparate phenomena, the numerical, the spatiotemporal and physical, and the linguistic, we notice two features. First, there are powerful reasons for believing that many significant features of these

---

[illegible] "At its most extreme, semantic, or meaning, holism is the doctrine that all of the inferential properties of an expression constitute its meaning"—Devitt, "A Critique of the Case for Semantic Holism," in J. Tomberlin, ed., *Philosophical Perspectives, vi: Language and Logic* (Atascadero, CA: Ridgeview, 1993).

[illegible] Similar relational notions are historically suggested in the slogan "meaning is use" that we see in views of J.L. Austin, the later Wittgenstein, and others. Relational views of beliefs are developed in works by Quine and J. Ullian in *The Web of Belief* (New York: Random House, 1970). Quine and Ullian do not seem to have a serious proposal for formalizing their "web," and they give a structure only for beliefs, not their composing concepts, according to certain entailment and theory-dependence relationships.

phenomena are extensively, or even completely, relational. This claim is strongly suggested by a deeper phenomenological investigation as well as by our most sophisticated theories of number, of space and time, of physical things, and of words, languages, and their features. Second, there is nevertheless a tendency to continue to speak, even among those who know these theories well, of "properties" of numbers, physical objects, and linguistic entities. I would even admit to an explanatory and computational (or psychological) bias in favor of such modes or presentation: Aristotelian and monadic modes of inference and conceptualization may be our default mode of our reasoning and naming. The world might also be so structured that such simplified—monadized—accounts cause comparatively little damage for everyday accuracy or usefulness; and computation itself might be such that dealing with relatedness in all of its glory (that is, in the full quantified, relational predicate calculus) might involve insuperable constraints on our ability to reach answers and effectively assess such structures—for example, Church's Theorem and the nondeterministic polynomial time (NP) completeness of many properties of relational structures.[17] Our minds and traditions may have opted for "reasonable" success in thought and speech, over mind-boggling and frequently ineffective efforts at accurate representation and manipulation of relational phenomena. This is not to be construed as admitting that we cannot ever conceive of, or talk clearly about, relatedness, but only that we often do not—and sometimes need not. We can, after all, think and talk about *some* mathematics, even if we often need the help of diagrams, and we easily use indexicals in natural language.

III. STRUCTURE, ASYMMETRIC GRAPHS, AND ARISTOTLE REFUTED

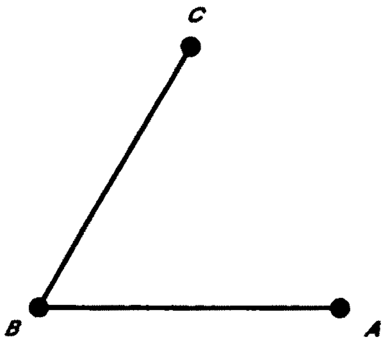
A graph is often defined set theoretically as a set of points, typically called "vertices" or nodes, and a set of "edges" connecting these points. The graph G can also be identified (misleadingly, as I shall soon argue) using set theory as an ordered pair consisting of the vertex set V and the edge set E, that is, $G = < V, E >$. If the graph is a "simple" graph, an edge is an unordered pair, a doubleton set consisting of two vertices. Such an entity may be considered to be the portrayal of the structure induced by a single two-place, symmetric relation. If the graph is "directed," or a digraph, then an edge is an ordered pair. A directed graph is a structure formed by a single two-place, asymmetic relation. We may also allow an edge to connect a vertex with itself, in which case a graph is reflexive. (Standard graphs are irreflexive: no edge joins the selfsame vertex.) Further extensions are possible. We might, for example, formulate a notion of edge that simultaneously connects three or more vertices: a hypergraph.

---

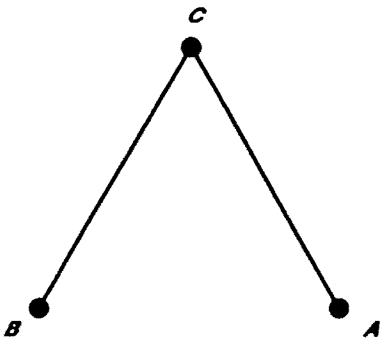
[17] More exactly, it is a combinatoric explosion that we see, for example, in the NP-completeness of the identification of graph isomorphism and of determining subgraph relations. See Michael R. Garey and David Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (New York: Freeman, 1979).

General graph theory is the theory of structures induced by a single relation (two- or more place, symmetric or not, and so on) and integrally involves what we might think of as “relational combinatorics”: the ways in which entities (here, vertices) may be connected by a relation. Observe, however, that it is the theory of how entities may be connected by just *one* relation, not many relations of diverse arities. As we shall see, even with this restriction, graphs embrace a high degree of complexity, diversity, and hence information. Even with small graphs—say, graphs with just twenty vertices—the diversity of structures that are in some sense distinct is dazzling and, indeed, largely uncontemplated.

Defining graphs in this set-theoretic way retains, however, some of the distinguishing flaws of normal logical and set-theoretical notation, namely, set-theoretic entities are distinguished even when we can see they have the same graph-theoretical structure. Two structurally identical graphs will be considered distinct if their individual vertices are different (or appear to be labeled differently). For example, consider two graphs, both with three vertices, *A*, *B*, *C*: graph 1 contains only the edges {*A*, *B*} and {*B*, *C*}, while graph 2 contains only the edges {*A*, *C*} and {*B*, *C*}. Since their respective edge sets {{*A*, *B*}, {*B*, *C*}} and {{*A*, *C*}, {*B*, *C*}} are different, then—if we *define* graphs as set-theoretic structures—graphs 1 and 2 are different. In a sense, however, the two have the *same* graph structure: they are just looked at, or labeled, in different ways.



Graph 1



Graph 2

Two graphs are structurally identical just when their labels can be rearranged so that they are set-theoretically identical. This relationship between two graphs is normally described as an isomorphism: two "distinct" graphs are structurally isomorphic just when there exists a 1-1 function, *f*, between the nodes of the first graph, and those of the second, that is edge preserving. Conversely, two graphs are structurally distinct if they are not isomorphic: if no function exists that maps vertices onto vertices in an edge-preserving way. We are primarily interested in structurally distinct graphs rather than in artifacts of the labeling of their vertices, or the way we have diagramed them or happen to be viewing them. The task of determining, visually or computationally, whether two representations of graphs are isomorphic (structurally identical) is nevertheless not a trivial one.

What we would ideally wish for the theory of pure graph structure is an account of graphs such that every difference between two graphs' descriptions entails a difference in structure, and vice versa. Existing set-theoretical and diagramatic notations do not do so because differences in vertex labels result in distinct (that is, non-set-theoretically equivalent) expressions, or in diagrams that look different but are not structurally distinct.[*] This obscurity arises because of the presumed "individual" vertex labels.

If we include quasi-distinctness—that is, distinct only in virtue of having differently-named vertices—the number of simple graphs is given by the formula $2^{(p(p-1))/2}$ where p is the order, or number of vertices (or rather, vertex names!). Even this combinatorial explosion is limited by the assumption that we are choosing from a canonically ordered set of vertex labels. For $p = 2$ vertices, this is 2; for $p = 3$, this is 8; for $p = 4$, 64; for $p = 5$, 1024; $p = 6$, 32,768; and so on. The number of nonisomorphic graphs initially rises much less steeply, demonstrating the nuisance that mere vertex naming causes. It is given by a difficult formula due to George Pólya,[*] and for increasing values of p starting with 2, generates the sequence:

---

[*] Since the graph isomorphism problem is NP-complete, there exists no canonical way of representing graphs so that we can always see when two graphs are isomorphic (or putting them into this canonical form is itself NP-complete).

[*] See N.L. Biggs, E.K. Lloyd, and R.J. Wilson, *Graph Theory, 1736-1963* (New York: Oxford, 1976).

Order (Number of vertices)	Number of nonisomorphic graphs
2	2
3	4
4	11
5	34
6	156
7	1,044
8	12,346
9	308,708[*]

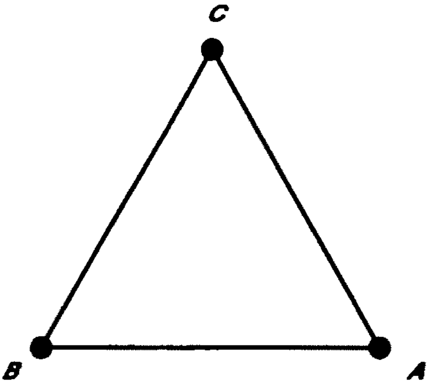
This sequence asymptotically approaches $2^{p/2}/p!$

Even the theory of "unlabeled" or nonisomorphic graphs has undesirable features for a theory of pure structure, I now want to argue. Here, I go beyond conventional graph theory and develop a way of looking at graphs for specific philosophical purposes that is not part of the usual mathematical theory of graphs (polluted as it has become by set-theoretic notational devices). The crucial concept for our purposes here is structural distinctness, free from all artifacts of labeling and perspective. The first phase of this investigation, discussed above, observed that mere aspects of vertex labels (or ways we draw or perceive diagrams) does not coincide with this interest. We home in on nonisomorphic graphs as paradigmatic of structural distinctness.

The second phase of this investigation attends further to the issue of vertex labels. Consider now only graph 1. Examine carefully the features of vertices *A*, *B*, and *C*, trying to ignore information misleadingly imported through vertex labeling. *B* is different—structurally different—from both vertex *A* and vertex *C*, namely, vertex *B* has *two* edges emanating from it: it is of degree 2. Both vertex *A* and vertex *C* have only one edge. In fact, on close examination, once we ignore their different labels, vertices *A* and *C* begin to fade into one another. They have the same structural features: both have only one edge emanating from them (they are of degree 1), and both are directly connected to a single vertex that has degree 2. Each is indirectly connected (by a two-step path through the unique vertex of degree 2) to the other. We might call the property that *A* and *C* have *graph-structural symmetry*: they occupy places in a graph that cannot be described, using only structural features, in any way that distinguish them. In the completely connected graph of order 3:

---

[*] F. Harary and E.M. Palmer, *Graphical Enumeration* (New York: Academic, 1973).



Graph 3

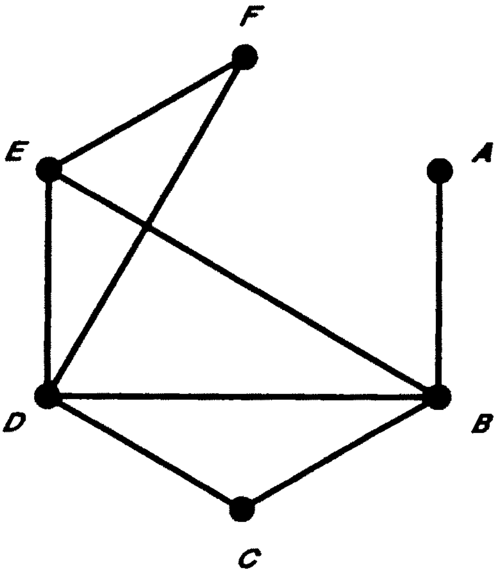
none of the vertices is structurally distinct: each is connected by two edges, and each is connected to every other vertex.

Observe that what I have called the two phases of our investigation of structural distinctness involve two different notions of distinctness. The first applies to graphs, and may be called *internal distinctness*: two entities (graphs) are distinct if and only if they contain features that structurally distinguish them. The second applies to vertices (and subgraphs) and may be called *external* (or *contextual*) *distinctness*: vertices have no internal structure, but are distinct if they are structurally related to other entities in a unique way. We do not rely upon *our* labelings to assert or assume they are distinct.

The exact criteria for distinctness and structure of such representations are of singular importance if we identify the concrete world, and its constituents, as graphs—or as any formal structure. We should not import assumed individuation of entities, whether they be logical individuals, graph labels, or proper names, when a key component of metaphysics is precisely the individuation of entities. A key metaphysical and methodological principle is that we shall not assert, or assume, metaphysical distinctness unless we can show (internal or external) structural distinctness. Traditional logical metaphysics has used the *deus ex machina* strategem for individuation (and thence for structure), while I propose lifting ourselves up by the bootstraps, using nothing but notions of mathematized structure. The method I am proposing for formal (not "logical") metaphysics has clear affinities with constructivism in the philosophy of mathematics.

A graph-structural symmetry between two vertices, X and Y, occurs when there is a 1-1 function, such that all nodes are mapped to themselves, except that X is mapped to Y, and Y to X, and that applying this function to the edge set of the original graph results in a graph that is isomorphic to the original. Other, more systematic symmetries may occur, even where there are no two nodes that are symmetric. The test for these broader (subgraph) symmetries will be whether or not there is a 1-1, nonidentity function, *f*, whose domain and range is the vertex-set, and which results (when applied to, or "translating," the edge-set) in a graph structurally identical to the original. If there is such a function, we shall say there is an auto-isomorphism, and more exactly that the graph is auto-isomorphic under *f*. Where there exists one or more auto-isomorphism, we shall say that the graph itself is symmetric, or contains a symmetry. This symmetry may consist either in containing one or more vertex symmetry, or in containing one or more subgraph symmetry, or both. A graph for which there is no such *f* will be called non-auto-isomorphic or, simply (following P. Erdős[31]), *asymmetric*: the graph contains no symmetries.

It is a surprising fact that there are no asymmetric graphs of orders 2, 3, 4, or 5; that is, all graphs of these orders contain some symmetries. We might then even suggest that these small graphs lack (nonexternally imposed) structural distinctness and thus are not well-defined structures. The first asymmetric graphs are of order 6, one of which is:



Asymmetric graph of order 6

---

[31] Erdős and A. Rènyi, "Asymmetric Graphs," *Acta Mathematica Academicae Scientiae Hungaricae*, XIV (1963): 293-315. See also M. Capabianco and J. Molluzzo, *Examples and Counterexamples in Graph Theory* (New York: North-Holland, 1978), pp. 89-101.

In an asymmetric graph, it is possible to give a unique, purely structural description for each vertex. For this asymmetric graph, one such description is:

A: Unique vertex of degree 1 D: Vertex of degree 4 not adjacent to A
B: Vertex [of degree 4] adjacent to *A* E: Unique vertex of degree 3
C: Vertex of degree 2 adjacent to two degree-4 vertices (*B*,*D*)
F: Vertex of degree 2 not adjacent to two degree-4 vertices

The literature on asymmetric graphs is extremely sparse. The number of structurally distinct, asymmetric graphs of order 6 is 7 (out of 156 nonisomorphic graphs). There are 152 structurally distinct, asymmetric graphs of order 7 (out of 1,044).[22] I do not have a guess concerning an enumeration theorem for asymmetric graphs, and there is no proof or speculation in the literature (to my knowledge): that is, how one calculates the number of structurally distinct asymmetric graphs for order *p*.

The theory of "pure" formal structure is for me the theory of nonisomorphic, asymmetric graphs, and their subgraphs. Graph theory, as it is usually investigated and described, contains too many artifacts of vertex naming. The study of graph theory through unlabeled diagrams is thus purer than that using conventional set-theoretical representations. Consequently, when I speak of what I propose is the conceivable structure of the world—the world graph for short—it is surely an asymmetric graph (if it is a graph at all). It is only asymmetric graphs that have structural distinctness that is not imposed from without, such as by naming.

Aristotle's principle refuted. The *locus classicus* for a discussion of relations in undoubtedly section 7 of Aristotle's *Categories*.[23] Aristotle there states a number of conjectures and arguments that have had a deep impact on the limited clear thinking about relations in Western philosophy. For instance, he assumes that no "primary substance" is relational, but broaches the possibility that some secondary substances are, such as "head [of]" or "hand [of]" (8a14f). At 8a37f, he begins an argument that no substance, primary or secondary, is relational. Apparently, then, Aristotle regarded relations just as ways of speaking that are ultimately reducible to monadic aspects of substance.

---

[22] I have explored these graphs and their properties in a computer-assisted way. I have written a large number of programs in Prolog that first identify nonisomorphic graphs, and then manipulate symmetries, asymmetries, and subgraph properties of these graphs.

[23] See a very fine article by Fabio Morales, "Relational Attributes in Aristotle," *Phronesis*, xxxix, 3 (1994): 255-74; Frank A. Lewis, *Substance and Predication in Aristotle* (New York: Cambridge, 1991); and J.R. Ackrill's *Aristotle's Categories and De Interpretatione* (New York: Oxford, 1963). Quotations are from J.D. Barnes, ed., *The Complete Works of Aristotle* (Princeton: University Press, 1984), whose translation of the *Categories* is by Ackrill.

His argument is as follows. Anytime we speak of an entity being related to another, we presume the distinctness of the relata—those things which are related. This distinctness cannot be further established by relations, since they would require, for their definition, the distinctness of *their* relata. Hence, relations require monadic properties in order to define their relata. His argument seems to be that irreducible relations are conceptually incoherent, or result in an infinite regress. Although Aristotle himself speaks rather tentatively of the implications of his arguments, most who have thoroughly contemplated the issue of the foundational nature of relational versus monadic properties have probably followed a similar pattern of reasoning. And it is this reasoning which has given the primary impetus toward what I have called *monadism*.

The main steps of this reasoning seem to be:

1. The definition, and an understanding, of relations requires the distinctness of relata.[34]
2. This distinctness of relata cannot be established except by monadic properties, on pain of incoherence or an infinite regress.[35]
3. So, monadic properties have a more fundamental metaphysical importance than relations, since relations depend on monadic properties, but not vice versa.[36]
4. Consequently, there are irreducible monadic properties, and all relational properties are reducible to monadic properties.

Thesis 1 seems to be the heart of Aristotle's anti-Platonism, here extended to relations: that relations are known through their extensions, and not "directly." We might call this *extensionalism*, or *particularism*, or something of the sort. I need not dispute it to make my point.

I assert that the existence of asymmetric graphs shows conclusively, for the first time in the history of philosophy, that thesis 2 is incorrect: namely, distinct relata—vertices in asymmetric graphs, for example—can be distinct (Aristotle: known definitely), and that this distinctness can arise through relations alone. With this evidence, I believe that most of the impetus toward monadism immediately collapses. (Another alleged argument against relationalism, a puzzle over asymmetric relations, is addressed below.) Indeed, I shall soon show that exclusive relationism is superior to exclusive monadism on at least one metaphilosophical criterion, a variant of Ockham's Razor. Furthermore, I believe that in section I, I have shown that Aristotle's assumption—that many or all "primary substances" are monadic—is also probably mistaken. Perhaps none is.

---

[34] "It is clear from this that if someone knows any relative definitely he will also know definitely that in relation to which it is spoken of" (8a37-38).

[35] This claim is explicitly attributed to Aristotle in Morales.

[36] "But as for a head or a hand or any such substance, it is possible to know it—what it itself is—definitely, without necessarily knowing definitely that in relation to which it is spoken of.... And if they [these relata] are not relatives, it would be true to say that no substance is a relative" (8b16-21).

IV. THE WORLD AS ASYMMETRIC GRAPH

The view I have proposed is a mathematical, formal, and structural metaphysics, but it is anti-atomistic and, in a precise sense, nonlogical. I would admit that I have not yet given a single, clear argument that the world is, or can only be conceived clearly as, an asymmetric graph induced by a single, two-place, symmetric relation. I have presented, first, what I believe are powerful considerations that most phenomena in the world and mind are ultimately relational. Second, I have presented a rarefied account of relational structure that avoids difficulties that I had pinpointed in logical metaphysics, notably, individuation by fiat. This account shows exclusive relationism, pace Aristotle and many others, to be conceptually coherent. In this last section, I would like to add considerations in favor of relationalism and exclusive relationalism, and to the view that one, symmetric, dyadic relation is necessary and sufficient for describing the structure of the world. I would also like to connect this extraordinarily abstract conception of the world, so distant from the way we speak and the way we believe we think, to more common-sense and scientific conceptions of the material and phenomenological world.

Relations not only can define basic entities, but they do so more efficiently and elegantly than do monadic properties. Using only monadic properties, it is necessary that there be, for example, four "primary" properties in order to define sixteen basic entities in the world. In general, we require a minimum of $\log_2 N$ monadic primary properties in order to distinguish N "basic entities"—such as ultimate particles (quarks) at a space-time point. The number of basic entities (N) in the universe is very large (we may suppose), so the number of primary properties necessary to distinguish them is also large. On the other hand, a single relation can be used to distinguish an arbitrarily large number of basic entities.[27]

---

[27] Namely, through the method of unique position in an asymmetric graph. This claim relies on the easily proven theorem that there exists at least one asymmetric graph for every order greater than 5.

This is, I believe, quite a powerful point. If one is an antirelationalist, that is, a monadist, then one must believe—if one believes there are many distinct entities in the world—that they are ultimately distinguished by the different combination of monadic properties each has. Far from being a minimal, Ockhamite assumption, this entails a commitment to a very large number of irreducible, exemplified, monadic properties. Reduction of arity (from 2 or higher to 1) of basic predicates in the world is thus possible only if one tolerates an explosion of the number of basic monadic predicates. This is artfully hidden from us by the casual assumption—in materialist metaphysics, for example—that entities are ultimately and simply distinguished by their locations in space, time, or space-time. For this account not to turn out to be disguised relationalism, however, we would require a commitment to a view of space (and time) as being constituted by unique one-place spatial properties: absolute locations—and a sufficient number of them to distinguish all basic material entities.[*] While it is possible that some such view is palatable, it seems wrongheaded to have such a view unwittingly forced on us by a commitment to a monadistic metaphysics without conscious regard to this commitment's far-reaching implications for other kinds of metaphysical inflation.

The theory of graph-theoretic structure is sufficient to account for all structure in thought or world. Minimally, it has the information-theoretic content to describe the complexity of the apparent world, it mirrors the "computational" difficulty we have in grasping this world, and it has the combinatoric texture to give a theoretically satisfying account of the nature of the world. That is, the world is of daunting size and complexity, parts of it are difficult precisely to isolate and conceive, but it is fundamentally made up of parts arranged in simple, graspable arrangements.

This is an extremely speculative assertion that a graph—large graphs anyway—have the same compositional "feel" as the world; and that the "facts" or sentences of first-order predicate logic of logico-metaphysical analysis do not. Could the universe be a graph—a particular asymmetric graph? The breaks into two questions. First, is it possible that the universe is just a formal structure, and that we can only think about the world *as* such a structure? Second, is it conceivable that the structure that is the world is just a *graph-theoretical* structure?[*]

---

[*] Even if one tolerates one basic space-time relation, little-discussed constraints must be imposed on it so that it ensures the individuation of all material entities. If the material structure is infinite—either spatially or temporally—this is extremely problematic, since there may be repetitions that defeat individuation. Even if it is finite, uniqueness is not guaranteed, except by some feature such as the asymmetries of graphs; but few have suggested that material entities can only arrange themselves in such asymmetric arrangements.

The graphical description of a world is an asymmetric graph. An asymmetric graph is a possible world, structurally described. An *entity* in such a world is any subgraph of a world graph. Such a subgraph is a subset of the vertices of the world graph, together with all of the edges in the world graph's edge set such that both vertices of each edge are in this subset. The entities closest to a common-sense notion of an entity will be *connected* subgraphs: subgraphs in which there exists a path through its edges from every vertex in the subgraph to every other vertex in the subgraph. A great many of these entities in the world are not themselves asymmetric graphs. Indeed, many of these subgraphs are isomorphic to other but ultimately distinguishable subgraphs.

I would like to make a claim that might seem preposterous: that even moderate-sized asymmetric graphs, on the order of forty to one hundred vertices, have compositional attributes like the perceivable world. Structures of similar sizes composed from monadic properties emphatically do not have these features. My guess would be that the world is a larger structure than one formed by one hundred connected vertices, but the point I want to make can be made about relational structures of even this "small" size. How many distinguishable entities would a world graph of this order contain? How many different asymmetric graphs—possible worlds—of this order are there? A graph of order forty contains 2^{40} subgraphs. (We shall consider the matter of their individuation and identification in a moment.) This is a very large number. Even if our own world contains a still greater number of basic particles in our best available account of microphysical structure, it is clear that even relatively small such graphs approximate the number of basic components in our universe (physically interpreted). One peculiarity of the graph-theoretic notion of an entity is that while there are 2^{40} distinguishable subgraphs, a great many of these subgraphs "overlap": they contain the same vertices and indeed may overlap in terms of graph structure, mutually containing one or more of the same further subgraphs. Far from being an undesirable oddity of graph-theoretic relational structure, I believe this phenomenon in fact mirrors difficulties in individuation versus apparent individuation in the actual world. The "two" electrons in the double-slit experiment used to explore quantum mechanics may be entities that share a great deal of metaphysical structure in precisely this sense, even if they are in other ways distinct. (This overlap of metaphysical structure plays the role of speculations about their wave-like nature, dropping their strictly particle-like, individual nature.)

---

[*] True, graph theory is at best the theory of *discrete* structure based on one symmetric, dyadic relation. Basing apparently continuous phenomena, such as space-time, in discrete structures such as the lattices of quark chromodynamics is becoming more common, although an argument for a full-fledged reduction is rarely made. Compare D.W. Kueker and M.C. Laskowski, "On Generic Structures," *Notre Dame Journal of Formal Logic*, XXXIII, 2 (Spring 1992): 175-83.

The information content of such a possible graph world can be roughly equated with the number of nonidentical entities of that size and kind. This amounts to the number of nonisomorphic graphs of order forty. F. Harary and E. M. Palmer (*op. cit.*, p. 242) give the number of order twenty-four graphs as about $1.95 \cdot 10^{57}$ and for order forty, the number is a great deal larger than that. In other words, there is a great deal of information content in a graph of that order.* These two points support my contention that a large graph is somewhat like the apparent gross morphology of the world itself in terms of number of distinguishable entities and in basic complexity or information content—what together we might call the world's metastructural texture. Indeed, these features are duplicable in strictly mathematical structures of a remarkably small absolute size, such as graphs of forty-one hundred vertices.

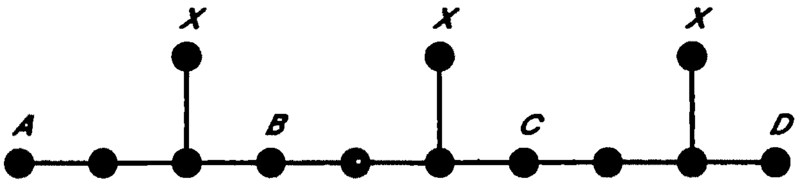
It is true that a great many of the "distinct" entities in the world graph, namely, many of its subgraphs, are in fact isomorphic. How then can such entities be said to be distinct? Subgraphs can be individuated in one of two ways. First, two subgraphs may have different internal structures: they are nonisomorphic. We might term this *internal* or *absolute* (*structural*) *individuation*. Second, they may occupy different "places" in the asymmetric world graph. We have built into the very nature of a world graph the fact that different locations, down to the level of vertices themselves, are distinguishable. Some subgraphs, while internally indiscernible, are individuated by their differing locations in the world graph. We may call this *external* or *contextual individuation*. Every subgraph in the world graph is individuated, either by its internal structure, by its "external" structure—or both.

A "property" of the entity identified with a subgraph is a *pattern*, either within that subgraph (an "internal" property) or in a larger subgraph (or subgraph skeleton) of which the given subgraph is a proper part (an "external" property).[41] A *relation* between two subgraphs (entities) is the pattern of a subgraph that connects the two given subgraphs or, less interestingly, is a formal relation that holds between descriptions of the two subgraphs considered in isolation. (The former is a type of a posteriori relation, that requires for its confirmation an examination of the world graph, not just the relata; the latter is an a priori or formal relation between the two relata.) A pattern is a subgraph skeleton: a subgraph in which all connected subgraphs of a certain specification are replaced with a single vertex. It is this feature of patterns/properties which will make the world seem to be organized hierarchically. There will be an extraordinarily large number of such possible patterns even in relatively small, connected graphs. This mirrors the plurality of distinct-seeming properties and relations in the world.

---

* Strictly speaking, we should be interested in the number of connected asymmetric graphs of order forty. I lack an enumeration theorem for asymmetric graphs, but would conjecture that they, too, exponentially increase and, in fact, asymptotically approach the number of nonisomorphic graphs for large orders. I owe this conjecture to conversation with graph theorist and my former colleague, H. Joseph Straight.

One respect in which the real world does not seem much like a graph is in this. In the real world, many and, in fact, the most interesting, relations are asymmetric, while in the world graph, the only basic relation is symmetric. How can so many asymmetric relations arise out of a symmetric relation, and why is the graph-theoretic account not more true to the phenomenal world? If entities are described as subgraphs, and relations are (extensionally) described as subgraph paths that connect these entities (at certain points), then a relation is asymmetric just when the subgraph path is asymmetric at the vertices where it joins its relata. (This is what we might call *internal-structural relational asymmetry*; there is a trivial sense in which all subgraph paths are externally asymmetric, since the points at which they join their *relata* are distinguishable in the asymmetric world graph.) The simplest such asymmetric subgraph path connecting entities *A*, *B*, *C*, and *D* will look like this:



Asymmetry via a symmetric relation

---

[41] In the physical or common-sense case, these would correspond to examples such as "having a flu virus" (internal) and "being in New Jersey" (external).

where *X* can either be a single node or a subgraph used to mark asymmetries.[*] A similar technique can be used to describe other features of relations, such as transitivity. The second question, of why we should not try harder to be more true to the apparent world, is far more difficult: it is in part the question of why one should use undirected rather than directed graphs—or of why one should not abandon graph-theoretic monorelationalism altogether. My best answer has to do with the desirability of seeking ultimate theories of metaphysical structure which are tractable and perfectly explain, and hence are mathematical accounts of discrete, probably finite, structures. Graph theory is far simpler, better understood, and more theoretically translucent than its structural competitors, including nominally the theory of directed graphs. (Either approach is far ahead of the polyadism of predicate logic.) The second reason is likely to be far less convincing: it is that symmetries have usually been regarded as the more convincing explanatory devices, and that asymmetries are symptoms of a missing variable or an ill-understood phenomenon. This may be a hint of a sophisticated methodological principle, or "merely" an aesthetic desideratum, but a related methodological principle has successfully driven modern theoretical physics. The suspicion of relationalism since Plato has arisen partly because of a mystery posed by asymmetric relations, for example: Which of two mutually entailing, converse relational facts, "*S* is shorter than *T*" or "*T* is taller than *S*," is more metaphysically basic? Neither seems metaphysically privileged, and yet it seems odd to admit both as basic facts when they are mutually entailing. One traditional resolution of this puzzle has been to use it to reject relationalism altogether and embrace monadism—for example, that each has a monadic property of height that grounds both relations. A less desperate response is to take this puzzle as evidence that all basic relational facts are symmetric, which do not allow the puzzle over relational converses to arise, and from which asymmetric relations may be derived.

I need to sketch how structuralism could plausibly analyze physical (and phenomenal) objects, their structure (what we might call *physical* structure), including redescribing space, time, fields and forces, and quanta. In short, how would my "theory of everything" (TOE) link up with theories of everything (TOEs) in contemporary physics?[43] My speculations are these. Physical objects, even the finest subatomic particles, certainly do not correspond to vertices. Instead, they themselves are composite entities, subgraphs of the world graph. Physical microstructure is graph-theoretic macrostructure.[44] Space and time are frameworks or grids that may conveniently, if sometimes imperfectly, be laid on the world graph.[45] If there is something that is an aspect of things or experience called "time," and if it is asymmetric, this is ultimately because there are asymmetric paths connecting certain subgraphs that are the entities and their states.[46] Forces, fields, and causal chains are also such paths among the subgraphs we identify as physical objects/states, and each is distinguished (if at all) by its structure. Some of the stranger physical phenomena are, I believe, accommodated without too much difficulty: the two electrons in the double slit experiment are not, for example, the fully distinct entities that we imagine them to be (or location and momentum are not independent properties). They graph theoretically overlap. Calling them *wave-like* is, I think, a less perspicuous metaphor than observing that metaphysical, graph-theoretic proximity or distinctness may not always perfectly correspond to spatial distance or distinctness, even for ostensibly physical entities.[47] Although physical structure is ultimately graph-theoretic structure if it is thinkable at all, this is not to say that we should always strive to analyze physical phenomena in terms of their ultimate graph-theoretic structure. All sciences and their goals of theoretical explanation aim for a certain level of useful description, and we might be perfectly satisfied with physicists' focus on levels of description that are no finer than useful for a theoretic account of "physical" objects.[48]

---

[*] A. P. Hazen later pointed out to me that D. Scott and M.O. Rabin had discovered a similar technique, described in Rabin, "A Simple Method for Undecidability Proofs and Some Applications," in Y. Bar-Hillel, ed., *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Conference* (New York: North-Holland, 1965), pp. 58-68, although not the general theory for the description of asymmetric relations in graphs.

---

[43] The views we see in a spate of recent works such as F. David Peat, *Superstrings and the Search for a Theory of Everything* (Chicago: Contemporary, 1988) and P.C.W. Davies and J. Brown, eds., *Superstrings: A Theory of Everything* (New York: Cambridge, 1988).

[44] Compare Armstrong: "It may be that at some deep level of analysis we come to particulars which are not spatio-temporal and spatio-temporally related," p. 90.

[45] Just as I have given what I think is a compelling argument why the world, if it is a graph, is an asymmetric graph, I also believe there are further such a priori restrictions on what the world, or the thinkable world, may be. I am not committed to precise Leibnizian claims about the "best" of all possible worlds, or about the plenitude (maximization) of "being" in the world, but I believe there are further such a priori, graph-theoretic reasons why the actual world graph is the single one that is. I believe (although as a regulative or "pragmatically justified" principle) that everything thinkable does have an ideal explanation and that all ideal explanation is ultimately a priori, indeed mathematical. There is one and only one asymmetric graph that has these features and constitutes the world, and thus only one alethically "possible" world or graph. Our sense of possibility and freedom is due to ignorance and computational intractability. For reasons that I do not understand, but suspect exist and are provable, there are in this unique graph (at least fairly) regular connecting paths among a certain subset of objects (subgraphs) that we conceive as mid-sized, phenomenal "physical objects," and that these paths have the structure we identity as space and time. Perhaps in very large, well-connected, asymmetric graph-theoretic structures, it is *always* possible to identify paths with certain loose specifications (transitivity, irreflexivity, and so on) that we associate with the somewhat vague, macrostructural notions of space and time.

[46] See Robin Le Poidevin, "Relationalism and Temporal Topology: Physics or Metaphysics?" *Philosophical Quarterly*, XL (October 1990): 419-32.

No theory of the ontological structure of the world could be considered complete unless it also connects to, and hopefully sheds light on, difficult issues in epistemology and the philosophy of mind. Here, I can only gesture toward possible conclusions, since my focus in this essay was of necessity on metaphysics rather than on the collateral topics of mind and thought. Thoughts are also subgraphs of certain sorts. Thoughts are about a phenomenon—refer to it—if they occupy a certain place in the world graph (in a "mind") and share certain structural features with the thought-about object: a thought, *per impossibile*, is perfectly about an object just when its internal graphical structure is the same as the object's, and when the object occupies a location in the system of all objects that is like the thought's location in the system of (that mind's) thought. This is an iconic view of semanticity which includes both internal and external iconicity.[49] Such a conception has

[47] What some metaphysics treat as "individual" discrete physical objects are really quasi-distinct, in appearing at some level (for example, to casual perception)—that is, with respect to some of their internal or external structure—to be metaphysically distinct. But they are not so, just as I share some DNA-patterns with my parents, my siblings, and beyond. More notoriously, the paradoxes of quantum mechanics crucially rely upon what I believe is the false assumption of complete metaphysical, physical distinctness of particles (in, for example, the double-slit experiment). I interpret quantum mechanics in fact, as a *reductio* of certain assumptions of metaphysical distinctness, or of the assumption of no action at a distance, or both. Similar views have led David Bohm, in *inter alia*, *Wholeness and the Implicate Order* (New York: Routledge, 1980), away from formalistic thinking altogether and toward a suggestive holistic Buddhism. (See Paul Teller, "Relational Holism and Quantum Mechanics," *British Journal for the Philosophy of Science*, XXXVII (1986): 71-81.)

[48] On the concept of nature, see "Natural and Artificial: The Fateful and Obscure Contrast," in my *Artifacts, Art Works, and Agency* (Philadelphia: Temple, 1993), chapter 12, and compare my holistic conception of nature with the suggestive views of Arne Naess (and some Eastern conceptions) in David Rothenburg, "No World but in Things: The Poetry of Naess's Concrete Contents," *Inquiry*, XXXIX, 2 (June 1996): 255-72.

[49] This unpopular view, and its place in the theory of meaning and reference, is described more extensively in my "Reflections on Iconicity, Representation, and Resemblance: Peirce's Theory of Signs, Goodman on Resemblance, and Modern Philosophies of Language and Mind," *Synthese*, CVI, 3 (March 1996): 373-97.

two advantages over current theories of reference and semanticity: first, it attempts to give a serious account of reference without appealing to the black box of denotation or to the *camera obscura* of causation, or scientistic "naturalized" or evolutionary epistemology. Second, it gives us a plausible handle on the many ways our concepts and thoughts can fail to be precisely about items in the world. That is, it gives an account of the types of defective thoughts that have intrigued but eluded modern philosophy from René Descartes to the latest theories of vagueness. Concepts are generally clearer the more we have of them: this is due to the necessity of external iconicity in disambiguating them. Even with respect to their internal structure, they may specify patterns (subgraph skeletons) that overshoot or undershoot their marks in various ways, that is, there are many, computationally distinguishable, ways in which thought subgraphs can be like—or fail to be like—their object subgraphs.

There might at first seem to be no place in these cold graphs for minds, consciousness, and other mental phenomena—unless, that is, everything is mental. Although within the dialectic of this essay it is wild and possibly irresponsible speculation, we should perhaps consider seriously the possibility that something like the pan-psychism of Spinoza, Leibniz, or Peirce is true, and that vertices are pure feelings (Peircean "firstnesses"), constituting a distinct thought or object only when connected to other such entities.

RANDALL R. DIPERT

United States Military Academy/West Point

REVIEW ESSAYS

Wittgenstein's Philosophy of Mathematics. PASQUALE FRASCOLLA. New York: Routledge, 1994. 208 p. Cloth $59.95; *Wittgenstein on Philosophy and Mathematics.* CHRISTOFFER GEFWERT. Åbo: Åbo Akademi University Press, 1994.

A comparison between these two books, published almost simultaneously, is extremely instructive. Pasquale Frascolla has written a very spare book. It aims to expound sympathetically, but hardly at all to evaluate, those of Ludwig Wittgenstein's philosophical views about mathematics which are expressed in writings of his that have been published. It therefore deliberately eschews any reference to those notebooks which, although accessible, have not yet appeared in print. It also eschews almost all discursive matter about his contacts or interaction with other philosophers or mathematicians, and severely sticks to expositions of the opinions he voiced at different stages of his career. Christoffer Gefwert's book is quite different in its treatment and its arrangement. It is crammed with collateral matter, describing Wittgenstein's career and his contacts with Bertrand Russell, Frank Ramsey, L. E. J. Brouwer, Friedrich Waismann, and others (though hardly anything is said about Gottlob Frege). The work of a devout disciple, it is concerned to state and defend Wittgenstein's views, and to contradict all interpreters not themselves committed disciples.

I. FRASCOLLA AND GEFWERT

Frascolla is well aware of the dangers of exegesis, and says in his Preface that it frequently happens that those fascinated by Wittgenstein's writings "remain, afterwards, prisoners of Wittgenstein's jargon and reformulate his ideas in such a way that there is no increase at all in our capacity of critical evaluation" (vii). This, indeed, is a danger run by all exegetes of philosophical texts, and Frascolla avoids it very well. Gefwert's book, by contrast, illustrates the danger to perfection. Frascolla aims, by careful examination of the texts, to arrive at a clear account of three phases of Wittgenstein's thought about mathematics; he does not paper over the difficulties inherent in it, but labors to extract a coherent resolution of them. Although his book will doubtless provoke angry expostulations from assorted Wittgenstein scholars to the effect that Wittgenstein did not mean this or that *at all*, he has succeeded extremely well; his book is a very valuable contribution to our understanding of some difficult and provocative ideas. Gefwert's aim is, rather, to controvert those whom

0022-362X/97/9407/359–374

Wittgenstein criticized and those who have rejected, or, in Gefwert's eyes, have misunderstood, his views; but his book is written with little understanding of either group, and fails to convey Wittgenstein's ideas clearly to those as yet unconverted. It therefore achieves little of value save to make salient certain confusions prevalent among a certain school of his followers.

Frascolla's book is divided into three chapters, dealing respectively with the *Tractatus Logico-Philosophicus*,[1] an intermediate period, and the later writings. By contrast, Gefwert deals with the *Tractatus* in his first chapter; then, after two historical chapters, he plunges into Wittgenstein's later views on mathematics, treating them as a unity and devoting the remaining five chapters to various aspects of them.

Unhappily, there is reason to doubt that this author, who, in solving the equation '$x^2 - 4 = 0$' (195), obtains only the positive root, is really competent to write about the philosophy of mathematics. Referring to Wittgenstein's notion of mathematical "prose"—that part of a mathematical demonstration that Wittgenstein felt entitled, as a philosopher, to criticize or reject—Gefwert cites as examples "the application of mathematical induction in Skolem's proof of the associative law of arithmetic" and "the notion of infinity occurring in Cantor's proof of the actual infinity of the transfinite cardinals" (139), and refers forward, for a fuller discussion, to his section VI.5 (actually VI.7). When the reader, perplexed to know what a proof of the actual infinity of the transfinite cardinals could be, turns to section VI.7, he finds, first, that Gefwert does not know what mathematical induction is, but confuses it with scientific induction in the sense in which this is contrasted to deduction. To explain mathematical induction, he quotes George Pólya as illustrating the use of induction, in the scientific sense, to arrive at a mathematical conjecture, a procedure he deprecates in Wittgenstein's name. Gefwert is here attempting to expound the very unsatisfactory passage in *Philosophical Remarks*,[2] §§ 163-65, in which Wittgenstein indeed says that the associative law cannot be stated, but only shown; but a denial of the legitimacy of basing a general statement about the natural numbers on mathematical induction demands a clearer argument than Wittgenstein, or Gefwert, supplies. In the pages on Georg Cantor, we are given much familiar material; but we never learn *which* proof of Cantor's it is that Gefwert refers to as "Cantor's so called proof of the 'actual infinite' " (226).

---

[1] D. F. Pears and B. F. McGuinness, trans. (New York: Humanities, 1961).

[2] R. Rhees, ed.; R. Hargreaves and R. White, trans. (Chicago: University Press, 1980).

The distressing truth is that Wittgenstein himself had in some degree the failing that his exegete here so egregiously exhibits. He read here and there in the logical and mathematical literature; but, eager as he was to exercise his superior insight, he seems frequently not to have got the hang of what he was commenting on. He failed to discover what those into whose works he dipped were trying to do; he was too certain that he had attained a viewpoint from which it could be seen as based upon illusion. Criticizing Thoralf Skolem, he wrote:

> The correct expression for the associative law is not a proposition, but precisely its "proof", which admittedly doesn't state the law, but shows it. And at this point it becomes clear that we cannot deny the law, since it doesn't figure in the form of a proposition at all (*Philosophical Remarks*, §165).

What is the point of lecturing *Skolem* in this fashion? Wittgenstein was criticizing the first proof in the celebrated paper of 1923 in which Skolem introduced primitive recursive arithmetic. In that system, there are no unbounded quantifiers; the associative law figures as a free-variable formula, which cannot be negated. You indeed cannot negate a law, but you can deny that it holds; in calling it a "proposition," Skolem required no more than that.

Gefwert assures us, frequently, that Wittgenstein denied the need for foundations, such as "Hilbert's idea of securing mathematics by...global (absolute) consistency proofs" (63). David Hilbert, however, was not seeking certainty, but *justification*, and that not for mathematics *in toto*, but for infinitistic procedures to which he could attach no direct meaning. If, like Hilbert, you think that general statements about natural numbers cannot be negated or subjected to other truth-functional operators, it is not unreasonable to suppose that reasoning that involves so subjecting them requires justification, necessarily indirect; nor to propose, with the intuitionists, that the logical constants be reinterpreted non-truth-functionally, and mathematics reconstructed on that basis. No amount of high-flown verbiage can make either procedure appear a priori absurd; only patient, sympathetic examination can show one or other to be misguided. What is bizarre is, rather, to deny that general statements express propositions that can be negated, and then to insist that mathematics, as practiced, requires no justification or amendment; and Gefwert succeeds no better than other commentators in removing the bizarre impression.

Gefwert devotes a whole section (VII.2) to Wittgenstein's opposition to Hilbert's program and his rejection of the distinction between object language and metalanguage, going on to discuss his views on contradiction and on Gödel's incompleteness theorems. He quotes remarks by Wittgenstein to the effect that "metamathematics" is "another calculus just like any other," and observes that, from his point of view, consistency "proofs are merely another mathematical system on the same level as the mathematical system to be investigated" (241). He charges a consistency proof for arithmetic with circularity, on the ground that it must employ induction, which is part of what is to be proved consistent (244). These comments are probably faithful to Wittgenstein's thoughts on the subject. Together with his polemic against the superstitious dread of contradiction, they reveal a misunderstanding of Hilbert's aims. The desired consistency proof would not, for him, be on the same level as the system to be proved consistent, for the latter involved infinitistic reasoning that the former would eschew. The consistency proof would indeed involve induction, but as applied only to free-variable statements, not ones involving quantification over all natural numbers. Proving the consistency of the system was not in itself the ultimate goal, but, by doing so, to obtain a finitistic guarantee of the correctness of finitistically interpretable conclusions of formal proofs.

Wittgenstein seems to have interpreted the notion of a metalanguage on the analogy of an attempt to survey language from outside language—a standpoint from which one could stand outside mathematics and guarantee the whole of it free from contradiction. Such an interpretation could be arrived at only by paying insufficient attention to what those he criticized had written. Of course, one who had no qualms about the meaningfulness of unbounded quantification over the natural numbers, or about its capacity to form statements with determinate truth values, would need no proof of the consistency of elementary arithmetic: What proof could be clearer than its intended model? Wittgenstein took no account of the qualms Hilbert had on these scores, nor of his acceptance of finitistic arithmetic as evidently valid. Gefwert tries to give sense to Wittgenstein's observations about Gödel's incompleteness result, as being to the effect that Gödel had not provided a *refutation* of Hilbert's program, but a reductio ad absurdum of it (260): it showed the absurdity of the attempt to prove by mathematics facts about mathematics. But the theorems are *about* formal systems: to object to them by rejecting the distinction between a formula within such a system and a statement about the system is to imitate the impertinent schoolboy who, when the teacher says, "Suppose we are given a formal system which...," objects, "But, sir, suppose we are *not* given a formal system."

One thing—surely not the only thing—that irked Wittgenstein about the first incompleteness theorem was the characterization of the undecidable sentence as true (not itself part of the content of the theorem, only of the "prose" Wittgenstein so greatly disliked). He never came close enough to discussing the actual content of the theorem to explain his objection to regarding as true a universally quantified sentence each of whose numerical instances is provable; but his attitude to mathematical truth is in any case puzzling. He apparently embraced Hilbert's "*non ignorabimus*": all mathematical questions can be decided. At the same time, there are no undecided mathematical propositions. A mathematical proposition is given sense by its proof: where there is yet no proof, there is not as yet a proposition. An interesting quotation from R. L. Goodstein (252) indicates that in 1935 Wittgenstein had rather a good understanding of the incompleteness theorem. He surely lost it. His way of thinking could greatly illuminate understanding; it could also destroy it, in the thinker as well as in others. The cavalier attitude that deprives much of Wittgenstein's writing on mathematics of real force is probably due to a fact that Gefwert emphasizes: his conviction, from 1930 onward, that philosophy, as he practiced it, was something absolutely new. He believed that the way he thought was utterly different from the way in which anybody had thought before; the thoughts of others could therefore be of value only as object lessons in how philosophical confusion arises. In this, he grossly exaggerated, not, indeed, his own acuity, but the novelty of the questions he asked and the answers he gave. Unhappily, some of his modern disciples have inherited this illusion; it is then, above all, that they are liable to make the error against which Frascolla warned.

Frascolla summarizes the view of mathematics presented in the *Tractatus* as comprising the following theses:

(1) The theory of classes is superfluous in mathematics.
(2) Frege's definition of the ancestral embodies a vicious circle. The correct way to express the 'and so on' is by means of an inductive definition.
(3) The natural numbers are exponents of operations, that is, indices to indicate how often an operation has been iterated, to be explained by means of another inductive definition. Numerical equations display the equivalence of indices differing in the manner in which they are symbolized.
(4) Equations therefore display formal relations: they therefore do not express propositions, but lack sense. They are not tautologies, be-

cause they are not constructed out of genuine linguistic expressions, but involve the equals sign, which attempts to state a formal relation that can only be shown.

(5) God would not need equations, but we do, because not all formal relations are perspicuous to us; the recognition of such a relation requires what may be called an intuition, arrived at by calculation and provided for by language.

(6) Mathematics is a logical method. In this sense, the Wittgenstein of the *Tractatus* may be said to have been a logicist.

(7) Although mathematics is not about symbols, the activity of the mathematician consists in manipulating symbols. In this sense, the Wittgenstein of the *Tractatus* may be said to have been a quasi-formalist.

Gefwert's chapter on the subject, though filled with collateral matter about Wittgenstein's relations with Russell and Ramsey, and a comparison of him to Immanuel Kant, gives a more cursory summary of the views expressed in the *Tractatus*, corresponding to (1), (3), and (4); it would be difficult for anyone who had not read the *Tractatus* to gain much impression of what those views were. Gefwert denies that Wittgenstein was in any sense a logicist, but this disagreement with Frascolla may be merely verbal. He quotes (35) the same passage as Frascolla (33) to demonstrate that Wittgenstein rejected the notion of intuition in mathematics altogether; Frascolla's more careful statement shows this up as a half-truth.

Frascolla gives a meticulous, even tedious, analysis of Wittgenstein's interpretation of the natural numbers, exposing a lacuna in his exposition; Gefwert distances himself much further from the text. Neither observes a significant point. It was an early result of Frege's *Die Grundlagen der Arithmetik: Eine logisch-mathematische Untersuchung über den Begriff der Zahl*[3] that G. W. Leibniz's proofs of numerical equations are defective in that they assume the associative law of addition without proof. Frege concluded that, even to prove numerical equations, we need some general laws. Gefwert bypasses this without comment by defining "the operation $x + 2$ by '$x + 2 = (x + 1) + 1$' " (39), which is unreasonable, since we need a *general* definition of addition, not a separate one for each number to be added. If the natural numbers are taken as the exponents of operations, however, the difficulty is genuinely bypassed, since associativity is built in from the outset, because the numbers act on the operation, but the operation on whatever is its base. The point is not of great weight; but if Frascolla had made explicit mention of it, his lengthy discussion would have become somewhat more interesting by offering some motivation for treating the numbers as exponents of operations.

---

[3] Breslau: Koebner, 1884.

The thesis underlying (4), that we cannot *say*, but only *show*, that a formal relation obtains, and, in general, that the a priori cannot be stated, but only shown, is the most famous one in the *Tractatus*; from it it follows that numerical equations, and mathematical statements generally, are only pseudopropositions. The thesis is the consequence of Wittgenstein's resolutely truth-conditional theory of meaning, which did not allow the consequence that sentences of logic and mathematics cannot be accorded senses such as those possessed by empirical ones to be evaded by appeal to such a device as Frege's *Sinn/Bedeutung* distinction.

Gefwert remarks (150) that Wittgenstein wrote about mathematics in two periods, 1929-34 and 1937-44, but Frascolla more perspicaciously assigns the year 1934 to the last of the three periods he discusses. In the second period, 1929-33, he says, Wittgenstein rejected the characterization of the natural numbers as exponents of operations, together with every attempt to define them or reduce them to something more basic: we need, rather, to explain the grammar of the numerals and of the word 'number'. Numbers are what numerals signify, but they are not objects of which the numerals are the proper names. Outside mathematics, a numeral is most characteristically used in a statement of number, which Frege rightly called a statement about a concept, of which the number is an external property. But 'concept' here must be restricted to empirical concepts. A number is an *internal* property of a mathematical concept to which it can be assigned; hence a statement of number *within* mathematics has a quite different meaning from an empirical one: the former is a proposition, the latter a grammatical rule. (By contrast, it was essential to Frege's theory that not merely physical objects, but objects of every sort, including numbers themselves, can be counted.)

Frascolla compares Wittgenstein's conception of elementary arithmetic to Hilbert's: for both, it concerns the manipulation of sequences of strokes, but Wittgenstein did not consider arithmetic to be *about* concrete signs, but as establishing *paradigms* or *standards* that determine grammatical rules governing descriptions of physical facts. Quoting some remarks by Wittgenstein opposing Frege's critique of formalism and drawing the inevitable analogy with chess (55), Frascolla observes acutely that these tally precisely with a passage quoted by Frege from the formalist Johannes Thomae, and wonders why Wittgenstein supposed he was differentiating his view from formalism. (There is something odd about Wittgenstein's dealings with the Frege/Thomae controversy. He wanted to repudiate the views of both disputants, but usually only made observations indistinguishable from those of one or of the other: much later, we find him writing: "It is their use outside mathematics, and thus the *meanings* of the signs, that makes the sign-game into mathematics"—an echo of Frege's "It is applicability alone that raises arithmetic from the rank of a game to that of a science.")

The distinctive feature of this second period, Frascolla says, is that Wittgenstein now allowed a notion of a mathematical *proposition*, which he understood in a verificationist sense. A sentence belonging to some class for which we know of a decision procedure would express such a proposition, whose sense we can grasp in advance of knowing whether it is correct or not: the rules for determining its truth value endow it with such a sense, and bear sufficient analogy to a procedure of empirical verification for it to be acknowledged as a proposition. When we do not know any such decision procedure, however, the form of words does not yet possess any definite sense: a proof, if discovered, will first give it one.

After quoting Wittgenstein's repudiation of the interpretation of the quantifiers as expressing logical product and sum, Frascolla wisely considers the more coherent discussion of Skolem's proof of the associative law from the *Philosophical Grammar*.[4] A suggestion that a universally quantified statement should be taken to *mean* that the induction basis and induction step hold good is nullified by Wittgenstein's rather arbitrary ruling that an inductive proof is not to be counted as a proof in the same sense as a sequence of algebraic transformations; it merely motivates our adoption of the formula so "proved" as a primitive rule of the system. (Frascolla does not enquire whether such a formula as '$1^3 + 2^3 + \ldots + n^3 = n^2 (n + 1)^2/4$' should also rank as a *primitive* rule.)

Frascolla remarks that this view of Wittgenstein's would not threaten Skolem's work, but that his views about real numbers appear more clearly revisionary. Wittgenstein rejected real numbers not generated by some effective process, say, of their decimal expansions, whether classically conceived or by means of free-choice sequences; he also rejected some that are so generated, such as the number whose binary expansion has 0 or 1 at its n-th place according as n is composite or prime. As for set theory, Wittgenstein rejected it altogether, save as a calculus whose interest is destroyed once its pretensions are recognized as illusory. Frascolla finds himself unable to defend Wittgenstein's claim that his comments touched only the "prose" surrounding the mathematical core.

---

[4] R. Rhees, ed.; A. J. P. Kenny, trans. (Berkeley: California UP, 1974).

With his third chapter, Frascolla confronts the dread *rule-following considerations* (RFCs), which he discusses more lucidly than do most who address the subject. He distinguishes, first, between an *empirical* statement that someone has followed a rule correctly and the *grammatical* statement that such-and-such is the result of following the rule correctly in a given case. The latter *defines* "correct application" for the case in question, and neither expresses a genuine proposition nor rests upon any facts, but provides the criterion for judging the truth of the empirical statement. The first component of the RFCs is the thesis that a rule cannot provide in advance for what is to be judged a correct application of it to a particular case; the connection between it and that application of it is made only when we recognize a particular result as constituting the correct application. Second, what gives the rule its meaning is not an interpretation, which could be formulated, but the existence of a practice of following it. This must be a social practice, since all that can distinguish following a rule from merely believing oneself to be following one is accordance with the community—not just with what its members do, but with what they acknowledge as being correct. Frascolla thus here endorses the "community view," in opposition to those, such as Colin McGinn, P. M. S. Hacker, and Gordon Baker, who maintain that the practice might be one engaged in by a single individual.

Frascolla emphasizes that the agreement between those who engage in the practice is not understood by us as *constituting* the correctness of an instance of it, but is, rather, a *presupposition* of our adopting the criterion by which it is judged correct. He criticizes Saul Kripke for running together the empirical statement that someone has applied a rule correctly with the grammatical statement giving the criterion for a correct application in that case; he thinks this is why Kripke had to invoke the distinction between truth conditions and assertibility conditions, whereas, given the criterion for, say, correctly adding 37 and 48, there is no problem about judging that someone has done so. This is unfair, however, since it ignores the fact that Kripke was expressly considering the case of an addition sum that had never been done before, so that no antecedent criterion for doing it correctly was available.

This matter is central to Gefwert's discussion of the RFCs. Before coming to it, he expresses his opposition to the *antirealist interpretation of Wittgenstein*. Kripke figures as a prime exponent of this interpretation: a deconstruction of his account of the RFCs is presented as the best way of coming to understand them, and, simultaneously, as showing what is wrong with the antirealist interpretation. This critique issues in a fantastic misunderstanding of Kripke:

> Each person who claims to be following a rule within a community...can be checked by others. But this requires that everyone involved in the practice of addition accepts the same *interpretation* of the rule of addition. It requires the competence to make an explicit decision concerning which rule is to count as the correct interpretation. The common interpretation, or disposition to follow the rule correctly, requires acceptance of the relevant assertibility conditions. It is the disposition to interpret these theoretical assertion rules in the same way which in the "community view" interpretation breathe [sic] life into a sign and what it signifies. According to Kripke's view we are thus led to believe that Wittgenstein sanctioned the insight that the expression "meaning is use"...is to be understood as the result of a correct interpretation where the rule (formulated as an assertion condition) is understood as a "theoretical entity" belonging to an *a priori* structure in the sense we see in the semantic "theories of meaning" of mathematics...(123; Gefwert's italics).

There follows a murky account of the RFCs, from which few will be able to grasp what Wittgenstein's contention was.

The polemic of Gefwert, as of others of like mind, against the antirealist interpretation rests upon a confusion between proposals for constructing a systematic theory of meaning and conceptions of how the meaning of an expression should be explained. It cannot be doubted that the later Wittgenstein was thoroughly opposed to the former; but no one who aimed, as he did, to explain the "grammar" of diverse expressions, and to do so by describing their use, could lack a conception of what such a description should consist in. In the *Tractatus*, he had a truth-conditional model for an explanation of sense. (Gefwert criticizes me (119) for interpreting the *Tractatus* as applying this model to mathematical statements as well as empirical ones. The context from which he quotes me might prompt such a misunderstanding; but, of course, I have never taken the *Tractatus* as advancing a Platonist view of mathematics.) The famous statement:

> In order to be able to say: "*p*" is true (or false), I must have determined under which circumstances I call "*p*" true, and I thereby determine the sense of the sentence (4.063).

was taken by the Vienna Circle to mean "I must have determined circumstances C such that in C I should say that 'p' was true," and hence interpreted as a verificationist dictum; but it was surely intended as "I must have determined circumstances C such that I should say that 'p' was true in C," and thus as endorsing a truth-conditional conception.

However that may be, it is certain that, in his later writings, Wittgenstein rejected that model. Frascolla quotes (118) a fatuous remark by McGinn that, given his adherence to a redundancy theory of truth, he could not deny that semantic sentences have truth conditions; the correct observation is that the redundancy theory precludes an explanation of the sense of a sentence by stating its truth conditions. The conditions we take as warranting the declarative utterance of a given form of sentence do not exhaust the use we make of it, but they form a central component of it, and must be included in any adequate description of that use, whereas the conditions under which, in general, a sentence of that form is true cannot be called features of its *use* at all, even if they could be characterized without circularity: they belong to a much higher theoretical level. There is really no room for disputing this: resistance to recognizing its truth arises from confusing it with the attribution to Wittgenstein of a systematic justificationist theory of meaning, which is, of course, out of the question.

The RFCs, understood as Frascolla understands them, have, as he stresses, a radical effect on our conception of mathematics. The notion of a mathematical proposition, grasped in advance of a proof (or refutation), sustained in the intermediate period for statements decidable by some algorithm, must now be rejected. What was regarded in that period as holding good for statements not so decidable, namely, that they acquired a definite sense only when a proof was found for them, now holds for the simplest formulas, such as numerical equations: one can grasp the sense only by knowing the proof and acknowledging it as correct. No description in general terms, given in advance, of a process of computation, let alone a framework of proofs, can constrain our recognition of a particular calculation as correct or of a sequence of deductive steps as valid. Acceptance of the calculation, or of the theorem, constitutes our adoption of a new rule of grammar, which excludes certain expressions as senseless. As Frascolla puts it, "the skill of the creative mathematician lies...in his ability to bring about a common inclination...to adopt a certain definition" (142); and the same applies to one who is merely skilled at carrying out calculations in accordance with established methods. This is indeed a radical form of conventionalism, as Frascolla acknowledges (135, 182n7). If the RFCs truly have the power to overturn the common-sense view that a mathematical statement may have a determinate sense before it has been proved or disproved, they are far from innocuous: either they are erroneous, or they demand a complete overhaul of almost everyone's way of thinking about mathematics.

Frascolla ends by discussing some detailed points, such as Wittgenstein's remarks about Cantor's proof of the nondenumerability of the reals and his attitude to transfinite cardinals generally, but wisely ignores his comments on Gödel. These discussions are far more sensible than those of Gefwert's; but it is a pity that even he fails to ask certain natural questions. Wittgenstein repeatedly asserts that it is only within mathematics that we can speak of there being denumerably many things of some kind. Would it then be evidently senseless to assert that there are denumerably many stars or galaxies? Again, we are told that, even before seeing Cantor's proof, the expression 'enumeration of all real numbers' has no clear sense for us; but it is not discussed whether we are worse off in respect of it than in respect of 'enumeration of all rational numbers' before the denumerability of the rationals is proved to us. Frascolla reviews Wittgenstein's dismissive remarks about the law of excluded middle, as applied to undecided mathematical statements, and has the same trouble as before in explaining his refusal, nevertheless, to propose revisions of mathematical practice.

Frascolla does not succeed in convincing us of the correctness of Wittgenstein's views on mathematics, and probably did not hope to; but he gives a careful, sympathetic account, not disguising difficulties and tensions, but doing his best by his subject and writing with the greatest possible clarity and the least possible mystification. His book is therefore of great help to anyone struggling to make out Wittgenstein's meaning; even those who think that he has mistaken that meaning ought to acknowledge this. Virtually the opposite is true of Gefwert. Except when he is recounting historical facts, his exposition is foggy; his conviction that Wittgenstein was not only right in everything he said, but had a wisdom of a superior grade to that of those whom he criticized or who criticize him lends to his writing a disagreeable tone and impedes him from thinking matters through.

II. THE PUBLISHERS' ROLE

Having treated of the work of the authors, I have now to discuss the work done, or, more exactly, neglected, by the publishers of these books.

Frascolla's book is written in a passable version of what may be called Academospeak, or, more politely, International English. It is highly creditable that a foreign academic should achieve so close an approximation to literate English: there are few indeed at British or American universities who could do half as well at composing a book in Italian or German. The book is not, however, in literate English. Words are used that do not exist in the language, such as 'ininfluent', 'resolutive', 'empiristic', and 'definitory'; the verb 'to result' is often employed, like Italian 'risultare', to mean 'to prove to be'; Frascolla says 'must not have been' when he means 'cannot have been', 'as so' when he means 'as such', 'believing to follow a rule' when he means 'believing that one is following a rule', 'monotonous' when he means 'monotonic', 'calling into question' when he means 'invoking', and 'that there are infinite prime numbers' when he means 'that there are infinitely many prime numbers'; he uses the wrong preposition, as in 'capacity of', 'inherent to', 'interest for', 'confer to' (for 'confer on') and 'collapse in' (for 'collapse into'); and he prefers noun clusters to adjectives or prepositions, as in the frequently occurring 'sign manipulation activity' in place of 'activity of manipulating signs' and 'grammar rule' in place of 'rule of grammar' or 'grammatical rule', this latter so often repeated as to become comic. Gefwert makes fewer bad mistakes, and even writes '[sic]' after 'principle' used as an adjective in a quotation from another writer of International English; but it *is* still International English that he writes. He uses 'substitute...with' (or '...for') to mean 'replace...by'; he says 'possibility to' when he means 'capacity (or ability) to'; he uses 'alternatively' when he means 'alternately'. He often has a plural verb after a singular subject, or conversely; and he uses 'due to' as a preposition in British Rail fashion, never employing 'due' as an adjective (since this occurs on almost every page, it becomes intensely irritating). He is very fond of the limp construction 'so far as...is concerned', which sometimes occurs twice in one sentence. Needless to say, Frascolla misses only one opportunity to split an infinitive, and Gefwert none (something that also occurs on almost every page). It is a definite *grammatical rule* of International English that an adverb or adverbial phrase must be placed after 'to' and before the verb root.

The blame is not the authors'; it is squarely that of their publishers. Joseph Conrad apart, every text not in the mother tongue of its author needs copy-editing; hardly anyone becomes sufficiently skilled in a language not his own to acquire a perfect ear for what will strike a native speaker as natural, or what will mean what he intends it to mean, or even for what is a genuine word of the language. From these books as published, one would suppose that neither publisher had any copy-editing done, but simply demanded from the author a text prepared on disk, to be slung straight off to the printer. The frequent occurrence in Gefwert's book of opening quotation marks as the last character of a line is internal evidence that this must indeed have happened to it. Routledge, on the other hand, has assured me that they did have Frascolla's book copy-edited. The copy-editing has been done so badly as to give an irresistible impression that none was done at all.

To omit copy-editing or to allow it to be done so cursorily as to leave numerous gross mistakes uncorrected is disgraceful: it is selling a product under false pretenses. With editors of newspapers and journals, book publishers have almost complete control over what specimens of the English language appear in print. In international journals, conference proceedings, and symposia, a great bulk of material written in International English appears every year, composed and edited by people whose mother tongue is not English and frequently deviating from the English of native speakers more widely than does Frascolla's book. So far as I know, no other language suffers in this way: and ours is at the same time under assault from those whose mother tongue *is* English, but do not know, or have learned to disregard, the rudiments of grammar. The health of our language has therefore become exceedingly precarious; and publishers of books in English-speaking countries have accordingly incurred a grave responsibility for preserving its health. Insouciance of the kind manifested in this instance by Routledge is a dereliction of duty that affects all of us who speak English and love it, and, indeed, generations yet unborn. Gefwert's book has a Finnish publisher, and it is understandable that an academic should wish to publish in English rather than Finnish; but foreign publishers, too, should have a care for the damage they are inflicting upon a language that is not theirs.

Copy-editing, which used to be done skillfully at all publishing houses, is now a dual menace. It is a menace, in these two cases, by either not being done at all or being done so carelessly as to create the effect of not having been done at all. I have been informed in writing by one well-known British publisher that they no longer undertake any copy-editing whatever. When it *is* done with any care, it is usually a menace of the opposite kind. Copy editors of a prevalent school no longer believe that the profession of writer exists: authors are, for them, mere suppliers of "copy" that it is their duty to transform into publishable prose, and so, in obedience to maxims of their own, they hack about every sentence of the text before them, producing a uniform style bearing no resemblance to that of the author. (One maxim, followed by the copy editor of a book of mine, was that one cannot write 'on the other hand' unless one has previously written 'on the one hand'.)

But the greater menace is word processing. When personal computers first became available, they appeared to most writers an unparalleled benefit: now they could handle sentences, paragraphs, and chapters as a sculptor handles clay. They have become a paradigmatic example of the rule that other people will quickly turn every labor-saving device into a labor-creating device. Publishers, who once accepted hand-written copy, crammed with deletions and insertions, now demand disks, if they do not actually demand camera-ready copy; sometimes they issue complicated instructions about the type of disk, the machine and word-processing program to be used, and even the typeface for which the text should be formatted. Ever more pressingly, the writer who hopes to be published has to acquire technical skills; meanwhile, his ability to compose decent English sentences matters less and less. The advantage to the publishers is manifest. They barely have to do anything but advertise and distribute the finished product, and then wait for the money to come in; they need do little with the text of the book save forward it to the printer. They feel secure, knowing that authors *want* their books to be published, and will probably put up with anything to achieve that end. But they are cheating the public and helping to mutilate the language. A small minority of publishers still both recognize their responsibilities and respect their authors; but unless an effective means is found to curb the irresponsible behavior of the majority, the language will be irreparably damaged, and the profession of writer will in truth be replaced by that of copy smith.

There is, of course, a reply to all this, namely, that English is no longer *our* language more than it is that of anyone else on the face of the globe. It is now an *international* language, a lingua franca or pidgin, or, in view of the existence of millions of unfortunate people whose mother tongue it is, a creole; anyone, no matter what his mother tongue, has as much authority over what is and is not English as do any of us. That *may* be how some writers of International English see the matter: it may be that it would no more occur to them to have the language of their articles checked by a native speaker than if they had written them in Latin, because they no longer think of English as a language that *has* native speakers. But, as a speaker of a once great language, I desperately hope that that is not how it is going to appear to people in general. Whether it does or not lies more in the hands of publishers than of any other group; whether they do anything to prevent it depends on how many of them care about anything but making money.

MICHAEL DUMMETT

University of Oxford

NOTES AND NEWS

The Editors regret the untimely death of Marx W. Wartofsky, Professor of Philosophy at Baruch College and the City University of New York Graduate Center, on March 4, 1997, at the age of 68. He earned B.A. and Ph.D. degrees in philosophy at Columbia University in 1948 and 1952. Wartofsky taught for 26 years at Boston University, starting in 1957, becoming Full Professor in 1967 and Department Chair the same year until 1973. In 1983, he was appointed Distinguished Professor at Baruch and the CUNY Graduate Center, serving several terms as Department Chair at Baruch. Wartofsky was the author of three major books: *Conceptual Foundations of Scientific Thought* (1968), *Feuerbach* (1977), and *Models: Representation and Scientific Understanding* (1979). He authored more than 80 articles on subjects ranging from Denis Diderot and Karl Marx to aesthetics and psychiatry. Between 1960 and 1987, Wartofsky and Robert S. Cohen co-edited the first 100 volumes of the Boston Studies in the Philosophy of Science. In 1976, he and Carol C. Gould co-edited *Women and Philosophy: Toward a Theory of Liberation*, the first American collection of feminist philosophy. For more than a quarter-century, Wartofsky was editor of *The Philosophical Forum*, a quarterly journal which he founded in 1970. Throughout his work ran a preoccupation with what he termed "historical epistemology"—the many ways in which social practices condition and affect the development of philosophy and the formulation of philosophical issues. At the time of his death, he was working on a book dealing with this subject. His wit, generosity, and intelligence will be missed by his friends and colleagues in the profession.

NEW BOOKS

MOONEY, EDWARD: *Selves in Discord and Resolve.* New York: Routledge, 1996. xvi, 140 p. Cloth $62.95, paper $17.95.

MORALES, MARIA M.: *Perfect Equality: John Stuart Mill on Well-Constituted Communities.* Lanham, MD: Rowman and Littlefield, 1996. xiv, 219 p. Cloth $57.50, paper $22.95.

MORGENSTERN, MIRA: *Rousseau and the Politics of Ambiguity: Self, Culture, and Society.* University Park: Penn State UP, 1996. xviii, 270 p. Cloth $40.00, paper $17.95.

NELKIN, NORTON: *Consciousness and the Origins of Thought.* New York: Cambridge, 1996. xv, 341 p. Cloth $59.95.

NINO, CARLOS SANTIAGO: *Radical Evil on Trial.* New Haven: Yale, 1996. xii, 220 p. Cloth $27.50; and *The Constitution of Deliberative Democracy.* New Haven: Yale, 1996. ix, 251 p. Cloth $30.00.

OLSON, ERIC T.: *The Human Animal: Personal Identity without Psychology.* New York: Oxford, 1997. x, 189 p. Cloth $29.95.

O'NEILL, ONORA: *Towards Justice and Virtue: A Constructive Account of Practical Reasoning.* New York: Cambridge, 1996. x, 230 p. Cloth $54.95, paper $18.95.

OSTERGAARD, SVEND: *The Mathematics of Meaning.* Aarhus: University Press, 1997. 222 p. Paper $27.00.

PERRY, MICHAEL J.: *Religion in Politics: Constitutional and Moral Perspectives.* New York: Oxford, 1997. viii, 168 p. Cloth $29.95.

PINK, THOMAS: *The Psychology of Freedom.* New York: Cambridge, 1996. x, 284 p. Cloth $49.95.

PRITCHARD, MICHAEL S.: *Reasonable Children: Moral Education and Moral Learning.* Lawrence: Kansas UP, 1996. xi, 199 p. Cloth $29.95, paper $12.95.

RACHELS, JAMES: *Can Ethics Provide Answers? and Other Essays in Moral Philosophy.* Lanham, MD: Rowman and Littlefield, 1997. xii, 245 p. Cloth $57.50, paper $19.95.

RADDER, HANS: *In and about the World: Philosophical Studies of Science and Technology.* Albany: SUNY, 1996. x, 225 p. Paper $21.95.

RADIN, MARGARET JANE: *Contested Commodities: The Trouble with Trade in Sex, Children, Body Parts, and Other Things.* Cambridge: Harvard, 1996. xiv, 279 p. Cloth $35.00.

REIMAN, JEFFREY: *Critical Moral Liberalism: Theory and Practice.* Lanham, MD: Rowman and Littlefield, 1997. xi, 277 p. Cloth $62.50, paper $22.95.

RESCHER, NICHOLAS: *Priceless Knowledge? Natural Science in Economic Perspective.* Lanham, MD: Rowman and Littlefield, 1996. xii, 188 p. Cloth $52.50, paper $21.95.

REY, GEORGES: *Contemporary Philosophy of Mind.* Cambridge: Blackwell, 1997. xv, 362 p. Cloth $54.95, paper $24.95.

RILEY, PATRICK: *Leibniz' Universal Jurisprudence: Justice as the Charity of the*

Wise. Cambridge: Harvard, 1996. xv, 338 p. Cloth $39.95.

ROCCO, CHRISTOPHER: *Tragedy and Enlightenment.* Berkeley: California UP, 1997. xv, 228 p. Cloth $40.00.

ROOCHNIK, DAVID: *Of Art and Wisdom: Plato's Understanding of Techne.* University Park: Penn State UP, 1996. xii, 300 p. Cloth $40.00.

ROOF, JUDITH: *Reproductions of Reproduction: Imaging Symbolic Change.* New York: Routledge, 1996. vii, 224 p. Cloth $55.00, paper $16.95.

ROSE, GILLIAN: *Mourning Becomes the Law: Philosophy and Representation.* New York: Cambridge, 1996. vii, 163 p. Cloth $49.95, paper $15.95.

RUSSELL, BERTRAND: *An Essay on the Foundations of Geometry.* New York: Routledge, 1996. xvi, 197 p. Cloth $59.95, paper $17.95; and *An Inquiry into Meaning and Truth.* New York: Routledge, 1996. xvii, 352 p. Paper $24.95; and *Mortals and Others: American Essays, 1931-1935.* New York: Routledge, 1996. iv, 176 p. Cloth $49.95, paper $16.95; and *Unpopular Essays.* New York: Routledge, 1996. xiv, 190 p. Paper $15.95.

SCARRE, GEOFFREY: *Utilitarianism.* New York: Routledge, 1996. viii, 225 p. Cloth $55.00, paper $17.95.

SCHECHTMAN, MARYA: *The Constitution of Selves.* Ithaca: Cornell, 1996. xi, 169 p. Cloth $25.00.

SCHMALTZ, TAD M.: *Malebranche's Theory of the Soul: A Cartesian Interpretation.* New York: Oxford, 1996. xi, 308 p. Cloth $55.00.

SCHRAG, CALVIN O.: *The Self after Postmodernity.* New Haven: Yale, 1997. xiv, 155 p. Cloth $22.50.

SCOTT, CHARLES E.: *On the Advantages and Disadvantages of Ethics and Politics.* Bloomington: Indiana UP, 1996. xii, 216 p. Cloth $39.95, paper $15.95.

SCRUTON, ROGER: *A Short History of Modern Philosophy: From Descartes to Wittgenstein.* New York: Routledge, 1995. ix, 302 p. Cloth $45.00, paper $10.95.

SHAPIN, STEVEN: *The Scientific Revolution.* Chicago: University Press, 1996. xiv, 218 p. Cloth $19.95.

SHARPLES, R.W.: *Stoics, Epicureans and Sceptics.* New York: Routledge, 1996. xiv, 154 p. Cloth $49.95, paper $16.95.

SHIVELY, RUTH L.: *Compromised Goods: A Realist Critique of Constructionist Politics.* Madison: Wisconsin UP, 1997. xi, 166 p. Cloth $27.95.

SIEGEL, HARVEY: *Rationality Redeemed?* New York: Routledge, 1996. xi, 234 p. Cloth $59.95, paper $17.95.

SINGER, IRVING: *The Creation of Value: Meaning in Life, Volume 1.* Baltimore: Johns Hopkins, 1996. xvii, 158 p. Cloth $25.00, paper $14.95; and *The Harmony of Nature and Spirit: Meaning of Life, Volume 3.* Baltimore: Johns Hopkins, 1996. xii, 222 p. Cloth $29.95.

SIORVANES, LUCAS: *Proclus: Neo-Platonic Philosophy and Science.* New Haven: Yale, 1997. xv, 340 p. Cloth $35.00.

SMITH, BRIAN CANTWELL: *On the Origin of Objects.* Cambridge: MIT, 1996. xii, 420 p. Cloth $35.00.

SMITH, STEVEN B.: *Spinoza, Liberalism, and the Question of Jewish Identity.*

New Haven: Yale, 1997. xvii, 270 p. Cloth $30.00.

SNIDERMAN, PAUL M., et al.: *The Clash of Rights: Liberty, Equality, and Legitimacy in Pluralist Democracy.* New Haven: Yale, 1997. xii, 291 p. Cloth $35.00, paper $18.00.

SPERBER, DAN: *Explaining Culture: A Naturalistic Approach.* Cambridge: Blackwell, 1996. Paper $18.95.

STAVROPOULOS, NICOS: *Objectivity in Law.* New York: Oxford, 1996. xii, 216 p. Cloth $65.00.

STECKER, ROBERT: *Artworks: Definition, Meaning, Value.* University Park: Penn State UP, 1997. xi, 322 p. Cloth $55.00, paper $18.95.

STOUT, ROWLAND: *Things That Happen Because They Should: A Teleological Approach to Action.* New York: Oxford, 1996. x, 189 p. Cloth $45.00.

SUMNER, L.W.: *Welfare, Happiness, and Ethics.* New York: Oxford, 1996. xii, 239 p. Cloth $35.00.

SUNSTEIN, CASS R.: *Free Markets and Social Justice.* New York: Oxford, 1997. vi, 407 p. Cloth $35.00.

SUPPES, PATRICK, and MARIO ZANOTTI: *Foundations of Probability with Applications.* New York: Cambridge, 1996. ix, 193 p. Cloth $59.95, paper $19.95.

THOMPSON, EVAN: *Colour Vision: A Study in Cognitive Science and the Philosophy of Perception.* New York: Routledge, 1996. xvi, 354 p. Cloth $65.00, paper $24.95.

TILES, MARY, and HANS OBERDIEK: *Living in a Technological Culture: Human Tools and Human Values.* New York: Routledge, 1996. xi, 212 p. Cloth $49.95, paper $16.95.

TONG, ROSEMARIE: *Feminist Approaches to Bioethics: Theoretical Reflections and Practical Applications.* Boulder: Westview, 1996. viii, 280 p. Cloth $65.00, paper $21.00.

TUCKER, KENNETH H.: *French Revolutionary Syndicalism and the Public Sphere.* New York: Cambridge, 1996. x, 284 p. Cloth $54.95.

VINTGES, KAREN: *Philosophy as Passion: The Thinking of Simone de Beauvoir.* Bloomington: Indiana UP, 1996. xii, 202 p. Cloth $39.95, Paper $19.95.

VISION, GERALD: *Problems of Vision: Rethinking the Causal Theory of Perception.* New York: Oxford, 1997. ix, 275 p. Cloth $45.00.

WALLACE, JAMES D.: *Ethical Norms, Particular Cases.* Ithaca: Cornell, 1996. xi, 171 p. Cloth $27.50.

WALSH, VIVIAN: *Rationality, Allocation, and Reproduction.* New York: Oxford, 1996. 304 p. Cloth $65.00.

WANG, HAO: *A Logical Journey: From Godel to Philosophy.* Cambridge: MIT, 1996. xiv, 391 p. Cloth $40.00.

WARDY, ROBERT: *The Birth of Rhetoric: Gorgias, Plato, and their Successors.* New York: Routledge, 1996. viii, 197 p. Cloth $55.00.

WASHBURN, PHIL: *Philosophical Dilemmas: Building a World View.* New York: Oxford, 1997. xxiv, 383 p. Paper $26.00.

WEST, DAVID: *An Introduction to Continental Philosophy.* Cambridge: Blackwell, 1996. viii, 278 p. Cloth $54.95, paper $21.95.

WIREDU, KWASI: *Cultural Universals and Particulars: An African Perspective.* Bloomington: Indiana UP, 1996. ix, 237 p. Cloth $35.00, paper $16.95.

ZAGZEBSKI, LINDA TRINKAUS: *Virtues of the Mind.* New York: Cambridge, 1996. xvi, 365 p. Cloth $64.95, paper $19.95.

NEW BOOKS: ANTHOLOGIES

ANTONACCIO, MARIA, and WILLIAM SCHWEIKER, eds.: *Iris Murdoch and the Search for Human Goodness.* Chicago: University Press, 1996. xx, 266 p. Cloth $48.00, paper $18.95.

ARRINGTON, ROBERT L., and JOHANN GLOCK, eds.: *Wittgenstein and Quine.* New York: Routledge, 1996. xvii, 283 p. Cloth $59.95.

BICCHIERI, CRISTINA, et al., eds.: *The Dynamics of Norms.* New York: Cambridge, 1996. ix, 222 p. Cloth $59.95.

BRANHAM, R. BRACHT, and MARIE-ODILE GOULET-CAZE, eds.: *The Cynics: The Cynic Movements in Antiquity and Its Legacy.* Berkeley: California UP, 1997. ix, 456 p. Cloth $55.00.

CARRUTHERS, PETER, and PETER K. SMITH, eds.: *Theories of Theories of Mind.* New York: Cambridge, 1996. xv, 390 p. Cloth $54.95, paper $19.95.

COTTINGHAM, JOHN, ed.: *Western Philosophy: An Anthology.* Cambridge: Blackwell, 1996. xxiii, 626 p. Cloth $74.95, paper $24.95.

CRANE, TIM, ed.: *Dispositions: A Debate.* New York: Routledge, 1996. viii, 197 p. Cloth $65.00.

DEFLEM, MATHIEU, ed.: *Habermas, Modernity, and Law.* Thousand Oaks, CA: Sage, 1996. 173 p. Cloth $65.00, paper $22.95.

DOHERTY, PATRICK, ed.: *Partiality, Modality, and Nonmonotonicity.* New York: Cambridge, 1996. x, 300 p. Cloth $64.95, paper $22.95.

DYZENHAUS, DAVID, and ARTHUR RIPSTEIN, eds.: *Law and Morality: Readings in Legal Philosophy.* Toronto: University Press, 1996. xii, 779 p. Cloth $80.00, paper $34.95.

ESTLUND, DAVID M., and MARTHA C. NUSSBAUM, eds.: *Sex, Preference, and Family: Essays on Law and Nature.* New York: Oxford, 1997. viii, 349 p. Cloth $35.00.

FACKENHEIM, EMIL L.: *Jewish Philosophers and Jewish Philosophy.* MICHAEL L. MORGAN, ed. Bloomington: Indiana UP, 1996. xviii, 269 p. Cloth $39.95.

FREDE, MICHAEL, and GISELA STRIKER, eds.: *Rationality in Greek Thought.* New York: Oxford, 1996. xv, 353 p. Cloth $60.00.

GEORGE, ROBERT P., ed.: *The Autonomy of Law: Essays on Legal Positivism.* New York: Oxford, 1996. xii, 339 p. Cloth $65.00.

GERSON, LLOYD P., ed.: *The Cambridge Companion to Plotinus.* New York: Cambridge, 1996. xiii, 462 p. Cloth $59.95, paper $18.95.

GOLOMB, JACOB, ed.: *Nietzsche and Jewish Culture.* New York : Routledge, 1997. xii, 282 p. Cloth $65.00, paper $18.95.

GOODIN, ROBERT E., ed.: *The Theory of Institutional Design.* New York: Cambridge, 1996. x, 288 p. Cloth $49.95.

GORDON, LEWIS R., ed.: *Existence in Black: An Anthology of Black Existential Philosophy.* New York: Routledge, 1997. xviii, 328 p. Cloth $59.95, paper $18.95.

HAHN, LEWIS EDWIN, ed.: *The Philosophy of Hans-Georg Gadamer.* Chicago: Open Court, 1997. xv, 619 p. Cloth $56.95; paper $29.95.

HART, W.D., ed.: *The Philosophy of Mathematics.* New York: Oxford, 1996. vi, 316 p. Cloth $55.00, paper $16.95.

HART, RICHARD, and DOUGLAS R. ANDERSON, eds.: *Philosophy in Experience: American Philosophy in Translation.* Bronx: Fordham, 1997. xi, 281 p. Cloth $30.00, paper $18.00.

HECKMAN, SUSAN J., ed.: *Feminist Interpretations of Michael Foucault.* University Park: Penn State UP, 1996. xi, 320 p. Cloth $45.00, paper $17.95.

HIRSCHMANN, NANCY J., and CRISTINE DI STEFANO: *Revisioning the Political: Feminist Reconstructions of Traditional Concepts in Western Political Theory.* Boulder: Westview, 1996. xiv, 281 p. Cloth $68.00, paper $22.00.

HOLLAND, NANCY J., ed.: *Feminist Interpretations of Jacques Derrida.* University Park: Penn State UP, 1997. xii, 249 p. Cloth $40.00, paper $16.95.

JONAS, HANS: *Mortality and Morality: A Search for the Good after Auschwitz.* LAWRENCE VOGEL, ed. Evanston, IL: Northwestern, 1996. xi, 218 p. Cloth $79.95; paper $22.50.

LEWIS, FRANK A., and ROBERT BOLTON, eds.: *Form, Matter, and Mixture in Aristotle.* Malden, MA: Blackwell, 1996. v, 290 p. Paper $21.95.

LIGHT, ANDREW, and ERIC KATZ, eds.: *Environmental Pragmatism.* New York: Routledge, 1996. xvi, 352 p. Cloth $65.00, paper $19.95.

LILLY, REGINALD, ed.: *The Ancients and the Moderns.* Bloomington: Indiana UP, 1996. xiv, 326 p. Cloth $49.95.

LOUDEN, ROBERT B., and PAUL SCHOLLMEIER, eds.: *The Greeks and Us: Essays in Honor of Arthur W.H. Adkins.* Chicago: University Press, 1997. x, 264 p. Cloth $48.00, paper $18.95.

LUCEY, KENNETH G., ed.: *On Knowing and the Known: Introductory Readings in Epistemology.* Amherst, MA: Prometheus, 1996. Paper $22.95.

MAN, PAUL DE: *Aesthetic Ideology.* ANDRZEJ WARMINSKI, ed. Minneapolis: Minnesota UP, 1996. v, 196 p. Cloth $49.95, paper $19.95.

MEYERS, DIANA TIETJENS, ed.: *Feminists Rethink the Self.* Boulder: Westview, 1997. ix, 278 p. Cloth $59.95, paper $20.00.

MOSER, PAUL K., ed.: *Empirical Knowledge: Readings in Contemporary Epistemology.* Lanham, MD: Rowman and Littlefield, 1996. vii, 468 p. Cloth $69.50, paper $24.95.

MOUFFE, CHANTAL, ed.: *Deconstruction and Pragmatism: Simon Critchley, Jacques Derrida, Ernesto Laclau, and Richard Rorty.* New York: Routledge, 1996. ix, 88 p. Cloth $49.95, paper $12.95.

O'HEAR, ANTHONY, ed.: *Karl Popper: Philosophy and Problems:* New York:

Cambridge, 1995. iv, 297 p. Paper $22.95.

O'DONOVAN-ANDERSON, MICHAEL, ed.: *The Incorporated Self: Interdisciplinary Perspectives of Embodiment.* Lanham, MD: Rowman and Littlefield, 1996. vii, 165 p. Cloth $52.20, paper $21.95.

O'NEILL, JOHN, ed.: *Freud and the Passions.* University Park: Penn State UP, 1996. 235 p. Cloth $40.00, paper $17.95.

PETERSON, GRETHE B., ed.: *The Tanner Lectures on Human Value, Volume XVII.* Salt Lake City: Utah UP, 1996. x, 402 p. Cloth $30.00.

POLANYI, MICHAEL: *Society, Economics, and Philosophy: Selected Papers.* R.T. ALLEN, ed. New Brunswick, NJ: Transaction, 1997. viii, 395 p. Cloth $44.95.

POPPER, KARL P.: *The Myth of the Framework: In Defense of Science and Rationality.* M.A. NOTTURNO, ed. New York: Routledge, 1996. xiv, 230 p. Paper $16.95.

POJMAN, LOUIS P., and ROBERT WESTMORELAND, eds.: *Equality: Selected Readings.* New York: Oxford, 1997. x, 325 p. Cloth $39.95, paper $18.95.

RUSSELL, BERTRAND: *A Fresh Look at Empiricism: The Collected Papers of Bertrand Russell, Volume X.* John G. Slater and Peter Kollner, eds. New York: Routledge, 1997. xxxix, 886 p. Cloth $185.00.

STEIN, STEPHEN, ed.: *Jonathan Edward's Writings: Text, Context, Interpretation.* Bloomington: Indiana UP, 1996. xix, 219 p. Cloth $39.95.

VELMANS, MAX, ed.: *The Science of Consciousness.* New York: Routledge, 1996. xii, 207 p. Cloth $65.00, paper $18.95.

WESTPHAL, JONATHAN, ed.: *Justice.* Indianapolis: Hackett, 1996. xxvi, 212 p. Cloth $27.95; paper $8.95.

SCHATZKI, THEODORE R.: *The Social and Political Body.* New York: Guilford, 1996. x, 228 p. Cloth $42.95, paper $18.95.

WARREN, KAREN J., and DUANE L. CADY, eds.: *Bringing Peace Home: Feminism, Violence, and Nature.* Bloomington: Indiana UP, 1996. xi, 235 p. Cloth $39.95, paper $18.95.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 8, AUGUST 1997

WHEN A PAIN IS NOT*

> Nature has placed mankind under the government of two sovereign masters, pain and pleasure.
>
> —Jeremy Bentham

Much to my surprise, a recent review article[1] in neuroscience began with the following quotation from Hilary Putnam: "The typical concerns of the Philosopher of Mind might be represented by three questions: (1) How do we know that other people have pains? (2) Are pains brain states? (3) What is the analysis of the concept pain?" Upon reflection, I decided that the quotation was entirely apt after all. Philosophers do take pain as their paradigm case in quite a number of projects, and the fact remains that we do not know exactly how pain processing works in the brain, so there is lots of room for wild and rampant speculation. A little digging into the philosophical literature uncovered a wide range of opinions and arguments regarding how we think about pain. Indeed, I daresay just about every conceivable position is currently held today by some leading thinker or other. (See table 1.) We find some philosophers and neurophysiologists arguing that pain is com-

---

* Earlier versions of this paper were presented to the Department of Philosophy, Virginia Polytechnic Institute and State University, in 1995, and at the Society for Philosophy and Psychology 1996 Annual Conference. My thanks to both audiences for their thoughtful and thought-provoking questions. I am most especially indebted to Patrick Croskery, Mark Gifford, George Graham, Marjorie Grene, Bob McCauly, and Harlan Miller, and to Daniel Dennett for his illuminating commentary.

[1] A.V. Apkarian, "Functional Imaging of Pain: New Insights regarding the Role of the Cerebral Cortex in Human Pain Perception," *Seminars in the Neurosciences*, vii (1995): 279-93.

0022-362X/97/9408/381–409

Eliminativist		Averill Churchland Dennett
Completely Objective	Intrinsic to part of body	Armstrong Newton Pritcher Wilkes
	Behaviorist	Wittgenstein
	Functional state	Davis Lycan Shoemaker
	Perception	Hall Tye
Relation		Graham Stephens Conee Nelkin Kaufman
Purely subjective	Mysterious	Kripke McGinn
	Nonmysterious	Gillet Grahek

Table 1. Possible philosophical positions regarding the nature of pain, with the names of a few of the more prominent adherents listed. See the footnotes for bibliographic information.[2]

pletely objective; it is either intrinsic to the injured body part, a functional state, a set of behavioral reactions, or a type of perception. We also find some philosophers and psychologists arguing that pain is completely subjective; it is either essentially private and completely mysterious or it does not correlate with any biological markers but is completely nonmysterious. Finally, we find a few philosophers who disagree with both conceptions and hold that pain is not a state at all; either it does not exist as we commonly conceive of it or it is an attitudinal relation. Furthermore, each of these positions has become grist for someone's mill in arguing either that pain is a paradigm instance of a consciousness state or that pain is a special case and should not be included in any general theory of consciousness.

---

[2] References for works not discussed elsewhere in this article follow: E.W. Averill, "Functionalism, the Absent Qualia Objection, and Eliminativism," *Southern Journal of Philosophy*, XXVIII (1990): 449-67; P.S. Churchland, "Consciousness: The Transmutation of a Concept," *Pacific Philosophical Quarterly*, LXIV (1983): 80-95; E. Conee, "A Defense of Pain," *Philosophical Studies*, XLVI (1984): 239-48; W.E. Dandy, *Bulletin from Johns Hopkins Hospital*, XLVIII (1933): 357-61; L. Davis, "Functionalism and Absent Qualia," *Philosophical Studies*, XLI (1982): 231-49; G. Graham and G.L. Stephens, "Are Qualia a Pain in the Neck for Functionalists?" *American Philosophical Quarterly*, XXII (1985): 73-80; R.J. Hall, "Are Pains Necessarily Unpleasant?" *Philosophy and Phenomenological Research*, XLIX (1989): 643-59; R. Kaufman, "Is the Concept of Pain Incoherent?" *Southern Journal of Philosophy*, XXIII (1985): 279-83; S. Kripke, "Identity and Necessity," in T. Honderich and M. Burnyeat, eds., *Philosophy as It Is* (New York: Penguin, 1979); W.G. Lycan, "Form, Function, and Feel," this JOURNAL, LXXVIII, 1 (January 1981): 24-50; N. Nelkin, "Pain and Pain Sensations," this JOURNAL, LXXXIII, 3 (March 1986): 129-48, and "Reconsidering Pain," *Philosophical Psychology*, VII (1994): 325-43; H. Putnam, *Renewing Philosophy* (Cambridge: Harvard,

We are left with several significant questions and puzzles. Among them are Putnam's. I do try to answer Putnam's three questions here, but I also want to do more than merely carve out my niche among the myriad of positions. I aim to offer a diagnosis for why we have so little agreement concerning the nature of our pain states. In brief, I believe that there are two reasons. First, for many philosophers, there is a basic failure to appreciate the fundamental complexity of our neuronal processing. This is the less interesting cause. My claim is that philosophers are enamored with dissociation experiments, but fail to understand their purpose, which is to individuate the component pieces of our larger systems. I argue that our pain sensory system functions according to the same basic rules of all of our sensory systems.

Second, for many psychologists, neurophysiologists, and philosophers alike, there is an explicit or an implicit reliance on some sort of gate-control theory of pain. Although theories of this ilk can account for several "low-level" puzzling cases involving pain (why it is that stimulating our nociceptors under certain conditions can alleviate pain instead of causing more, for instance), they are notoriously vague when it comes to discussing the central gating mechanisms. This vagueness, I believe, obscures the fact that we actually have two separate systems involved in our perceptions of pain. One functions as a *pain sensory system* (PSS), quite analogous to our other sensory systems. The other—*pain inhibitory system* (PIS), which developed independently of our PSS—actively inhibits its functioning.

Differentiating between the two systems helps explain the remaining controversies surrounding the basic nature of pain. While a PSS

1992); S. Shoemaker, "Functionalism and Qualia," *Philosophical Studies*, xxvii (1975): 291-315, "Phenomenal Similarity," *Critica*, vii (1975): 3-37, and "Absent Qualia Are Impossible," *Philosophical Review*, xc (1981): 581-99; M. Tye, "A Representational Theory of Pains and Their Phenomenal Character," *Philosophical Perspectives*, ix (Atascadero, CA: Ridgeview, 1995), and *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind* (Cambridge: MIT, 1995); K. Wilkes, *Physicalism* (London: Routledge, 1977).

supports a perceptual view of pain as a completely objective phenomenon, adding in a PIS (without explicitly recognizing that fact) accounts for the strongly subjective aspects of pain. I shall claim that a PSS functions according to the same basic rules of all of our sensory systems and that, insofar as the pain system is a simpler system than, say, vision or audition, it makes sense to take pain perception as a paradigm instance of a conscious experience. But insofar as we also have a PIS, pain also becomes a special case in our collection of conscious phenomena. Hence, contra Putnam, we should not be using the experience of pain as an intuitive and unproblematic example of consciousness.

As a final conclusion emerging from my understanding of pain, however, we shall see that the sensation of pain—what most philosophers of mind focus upon as absolutely central to being in pain—is neither a particularly fundamental nor a particularly important component of our pain processing. One current popular research question in the philosophy of mind is determining whether some philosophical approach or other (for example, identity theory, functionalism, weak supervenience) can capture in an appropriate way what sensations *feel like*. If I am right about how we should understand pain, then the fervor devoted to this project might better be spent elsewhere, for what something is like becomes less important in explaining our mind.

I. THE COMPLEXITY OF OUR SENSORY SYSTEMS

Let me begin by outlining a few facts regarding our other sensory systems. I do this as a preliminary to discussing pain not because we understand, say, visual processing, so much better than pain processing—we do not—but because many of the facts of perceptual processing regarded as commonplace (even among philosophers of mind) are the same sort of facts that seem to confuse philosophers and psychologists when theorizing about pain.

Our visual system is quite complex, spans many areas in the brain, and is comprised of several subsystems whose interactions remain a mystery. It is widely known that different aspects of visual processing occur in different processing streams. For example, color is processed in intralaminar pathway, while motion is processed in the magnocellular. The auditory system works in an analogous fashion (though the interactions of its subsystems are not as mysterious). The medial superior olive of our auditory system probably computes sound location using interaural time differences. The lateral superior olive, on the other hand, computes sound location by using differences in interaural frequency.

What is important to notice is that it is quite all right for there to exist more than one processing stream in each modality. We might be mystified how color gets joined with shape and motion so that we have unified visual experiences of particular objects.[3] But we are not confused about whether the neuronal paths involved in computing an object's color are visual, or whether computing interaural time differences is auditory. We are perfectly happy to have each modality be involved in several maybe ultimately unrelated computations. We say (or, at least, I say) the parts of the brain that normally respond to impinging photons are part of the visual system, and the parts of the brain normally sensitive to air compression trains are part of the auditory system.

Naturally, this is a gross oversimplification of how our sensory modalities are actually individuated: without unpacking what is meant by "normal functioning," the definitions are virtually unworkable. By way of partially rectifying this gloss, let me briefly touch upon the top-down and bottom-up investigative methodologies in neuroscience (and in psychology, to some degree), for these analytic tools help disambiguate what counts as normal functioning. More importantly, they allow us to make claims about which computational algorithms and cell assemblies are and are not included in our brain systems and subsystems.

First, scientists use the method of *double dissociation* to isolate the processing streams that comprise our subsystems. If we can get *X* to occur without *Y* and also *Y* to occur without *X*, then scientists take this as grounds to claim that *X* and *Y* function as independent units. For example, explicit priming tasks in psychology demonstrate that we can record the meaning of a word or phrase without storing its syntax; implicit priming tasks show that the syntax of word or phrase can influence later linguistic processing while the meaning remains inert. I call this a *top-down strategy* because we start with a crude parsing of our system writ large (for example, linguistic processing) and then divide that system into its component pieces (syntactic processing, semantics). This method of investigation forms the backbone of Daniel C. Dennett's[4] functional decomposition.

Second, scientists rely on a *teleological analysis* to unite the various and sundry parts into wholes. Breaking down larger pieces into smaller ones is not enough to get the explanatory job done, especially when several of our systems overlap inside the head. Our brain houses lots of individual processors; knowing all the pieces does not identify the larger puzzles. Why do scientists believe that color and motion processing belong to the same system but that echolocation belongs to something else? This is not a trivial question since each of the subsystems is (mainly) anatomically and physiologically distinct from the other, and since individual neurons do not know the sort of signal to which they are responding. The information contained in an atmospheric compression wave or a photon wave triplet is transmitted as electrical and chemical energy once one moves inside the body.

---

[3] This is known as the binding problem in psychology; for discussion, see my "Psychology's Binding Problem and Possible Neurobiological Solutions," *Journal of Consciousness Studies*, I (1994): 66-90.

[4] *The Intentional Stance* (Cambridge: MIT, 1987).

Scientists use three converging strategies to isolate and construct systems from the component dissociable subsystems. First, they look for correlations between neural firing patterns and events in the external world, very much what Fred Dretske[5] has in mind with his informational semantics. Neurophysiologists take the smallest pieces of the puzzles, usually individual neurons, or the extracellular spaces around small groups of neurons, and record what they do under a variety of circumstances. They conclude that our color and shape detectors belong together because they are active under similar circumstances, namely, when the organism's retinas are bombarded by photons. Auditory cells are active in different contexts. Luckily for the scientists and their correlation project, true polymodal cells are relatively rare.

Second, scientists look at the neural connections fore and aft. Aside from knowing how a cell resonates with the environment, they also need to know to what this cell is connected—where the information the cell lights up to goes—and what is connected to the cell—how it gets the information to which it does respond. Determining the processing algorithm of any cell group is not as easy as it might sound; it is not a matter of merely recording all the stimuli it likes and then deciding what all the stimuli have in common. As S. R. Lehky and T. R. Sejnowski[6] remind us, even cells that we think we know well, such as D. H. Hubel and T. N. Weisel's[7] simple edge detectors, might not be involved in the computations we think they are. It is entirely likely, given Lehky and Sejnowski's simulation results, that the so-called "edge-detector" cells actually are involved in computing an object's axes of curvature.

---

[5] *Knowledge and the Flow of Information* (Cambridge: MIT, 1981), and *Explaining Behavior: Reasons in a World of Causes* (Cambridge: MIT, 1989).

[6] "Network Model of Shape-from-Shading: Neuron Function Arises from Both Receptive and Projective Fields," *Nature*, cccxxxiii (1988): 452-54.

[7] "Functional Architecture of Macaque Monkey Visual Cortex," *Proceedings of the Royal Society of London B*, cxcviii (1977): 1-59.

For another, more striking example, consider synesthesia, a condition in which one gets a bimodal experience from monomodal inputs—one can see and hear colors or taste and hear words.[8] Although we would probably want to say that someone seeing blue columns when she hears a bell ringing is having a visual experience, we would not want to say that atmospheric compression waves are visual inputs, even for this person. The inputs are still auditory; they come in through the ears and pass through the traditional auditory centers. They just happen also to travel through some of the visual pathways (probably via the limbic system). If we only had access to single-cell recordings of synesthetic cells, we would obviously misidentify what those cells were doing. Knowing how things are connected prevents us from leaping to what would otherwise be an entirely rational (but also entirely false) conclusion.

Finally, scientists consider historical and evolutionary facts whenever possible. We are biological organisms equipped to move through our environment. We evolved that way because (roughly speaking) those who can move most effectively through their environment succeed in reproducing the most. When thinking about our perceptual systems, especially when worrying about various components' purposes, it is important to keep in mind how the hypothesized system or subsystem is supposed to function with regard to motor assembling. For most, if not all, information processing in the brain is related to the motor system in one way or another. For example, the visual areas all have at least some indirect contact with some motor structure or other, either the basal ganglia, or the motor cortex, or the tectum, or something.[9] Motor information needs to be "siphoned off" the visual pathways at all stages along the ascending route so that the visual input can be used for motor output.[10] Quite often what seems strange or curious from a psychological point of view seems quite natural from an evolutionary standpoint.

If we can group subsystems together into larger systems via their function—which is just what it is about that system that increases the reproductive rate of the organism that houses it[11]—then so much the better. The brain puts great emphasis on the priority of motor tasks, and we should pay attention to this emphasis. Whatever purpose we ultimately propose has to fit with our biological natures. (Often, however, such considerations are not possible or are little better than just-so stories, for the details of the advantages have been lost over evolutionary time. Why do we see in color, for example? What reproductive advantage would it have given our ancestors long ago? The answer is not easy, nor is it clear.)

---

[8] R.E. Cytowic, *Synesthesia: A Union of the Senses* (New York: Springer, 1989), and *The Man Who Tasted Shapes: A Bizarre Medical Mystery Offers Revolutionary Insights into Reasoning, Emotion, and Consciousness* (New York: Putnam, 1993).

[9] E.R. Kandel and J.H. Schwartz, *Principles of Neural Science* (New York: Elsevier, 1985, 2nd ed.).

[10] P.S. Churchland, "Epistemology in the Age of Neuroscience," this JOURNAL, LXXXIV, 10 (October 1987): 544-53.

[11] Cf. C. Wright, "Functions," *Philosophical Review*, LXXXII (1973): 139-82.

I call this collection of research strategies *bottom up* because we begin with the smallest units in the brain and then arrange them into nested hierarchies. Based on gross similarities in response patterns, connections to other systems and organs, and putative selective advantages, we group the double-dissociated subsystems into hierarchically arranged classes. The process is not cut and dried by any means, but it is the best we have at the moment. Perhaps someday we shall be able to identify definitively to which system various cell assemblies belong in virtue of the inherent rhythms of the cells' firing patterns, or something like this.[12] Until then, though, the best we can do is to make educated guesses based on converging evidence.

Both approaches are required for a complete explanation of psychobiological phenomena. By breaking cognitive engines into interacting component pieces, the top-down strategy helps explain why organisms behave the way they do; and by categorizing and grouping the isolated parts, the bottom-up strategy helps explain what purpose the analyzed behavior serves. Reminding ourselves that we use both strategies in understanding our neural systems will rid us of the tendency to make our pain system into a cartoon; and reminding ourselves of biological heritage will aid in justifying a counterintuitive system that prevents our pains from occurring. Ultimately, I claim that our system for perceiving pain works in exactly the same fashion as our visual and auditory systems: it is a complex system with dissociable subsystems. Furthermore, it is a system that appears quite natural when considered against an evolutionary backdrop.

II. A SKETCH OF OUR PAIN SYSTEM

The classic view of our basic pain system is of two three-neuron subsystems.[13] (See figure 1.)[14] Each subsystem has a set of neurons which

---

[12] R. Emmers describes modality specific firing patterns in the thalamus in *Pain: A Spike-Interval Coded Message in the Brain* (New York: Raven, 1981).

[13] Even the classic story is becoming more complicated; see E.R. Kandel, J.H. Schwartz, and J.J. Jessel, *Principles of Neural Science* (New York: Elsevier, 1995, 3rd ed.), chapter 27.

[14] I owe this vision of the classic view to S.A. Cross's excellent review article, "Pathophysiology of Pain," *Mayo Clinic Proceedings*, LXIX (1994): 375-83. See also Kandel and Schwartz; and P. Roland, "Cortical Representation of Pain," *Trends in Neurosciences*, XV (1992): 3-5.

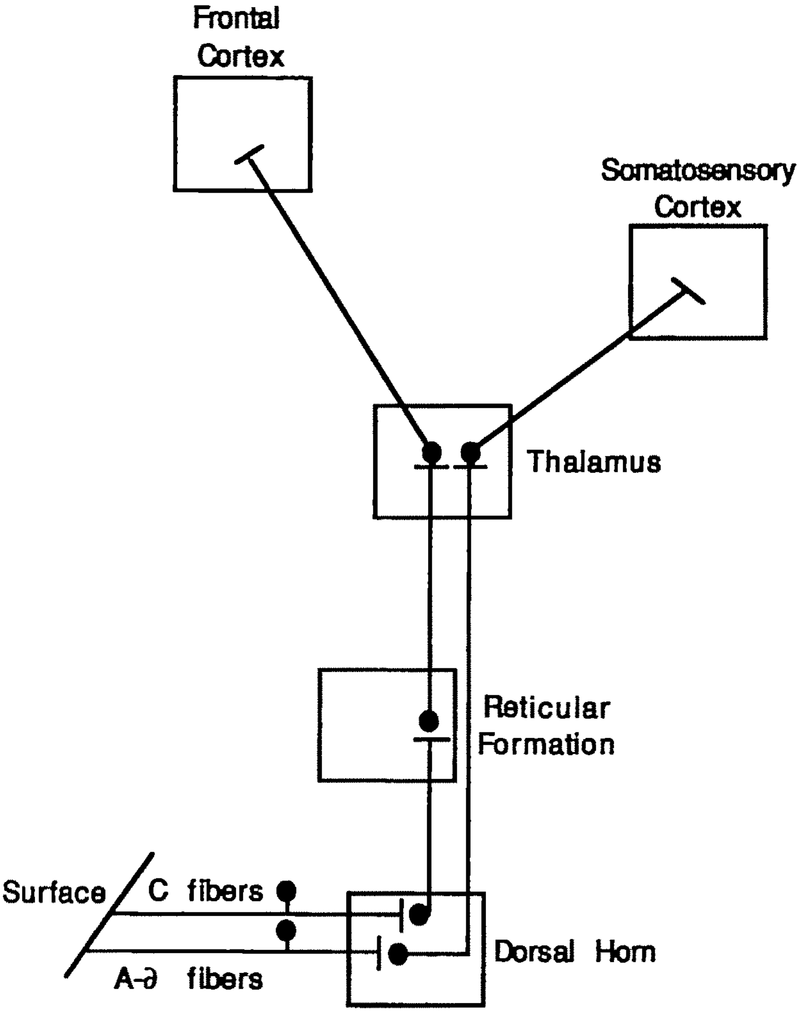


Figure 1. Diagram showing our pain sensory system. The first set of neurons take in information from the periphery and then synapse with a second set of neurons in the dorsal horn. These neurons ascend, some terminating in the reticular formation of the brain stem, others traveling to the thalamus. Axons that terminate medially in the thalamus synapse with a third set of neurons that project to the frontal cortex. Those which terminate laterally synapse with neurons that project to the somatosensory cortex.

resides in the dorsal root ganglion of the spinal column. These neurons extend their axons to whatever tissue they innervate and receive external input there. They also have a second axon that projects across to the dorsal horn. The axon in the dorsal horn connects with a second set of neurons housed in the dorsal horn whose axons run out of the spinal column and up to the thalamus. The third set of neurons projects from the thalamus to the postcentral gyrus in the cerebral cortex.

In 1911, H. Head and G. Holmes[15] proposed a dual system of afferent projections in our pain sensory system: an *epicritic* system that processes information regarding intensity and precise location, and a *protopathic* system that delivers the actual pain sensations. Eighty-six years later, we still believe they were fundamentally correct. We now know that we have a "sensory discriminative" subsystem that computes the location, intensity, duration, and nature (stabbing, burning, prickling) of the stimuli. This subsystem is subserved by the A-∂ fibers. These mechanoreceptive neurons are mylinated, so information can travel quite quickly along them (approximately 5-30 m/sec, as opposed to .5-2 m/sec for information traveling along unmylinated pathways[16]). Consequently, they transmit what is known as "first pain" or "fast pain." The threshold for activation is constant from person to person (threshold here), and this subsystem remains active (assuming no other defects in the organism) only as long as the raw nerve endings are stimulated.

We also have an "affective-motivational" subsystem that supports the unpleasant part of painful sensations. This system feeds directly into our motor response systems and is considered to be phylogenetically older than other aspects of our multifaceted pain system. This polymodal subsystem begins with the well-known unmylinated C fibers. Once they are activated, they will continue to fire for some time, even after the noxious event has ceased. This subsystem gives rise to what is known as "slow pain" or "second pain," so-called because this is what we feel second whenever we are injured—a diffuse and persistent burning pain. When someone has chronic pain, a protracted second pain is what is being referred to.

Similar to the color and form processors in the visual system, the A-∂-fiber and C-fiber pathways remain largely segregated. For example, generally speaking, they terminate in different layers on the dorsal horn. But there is more interaction than what we find in either the visual or auditory system. The dorsal horn contains *wide-dynamic range* (WDR) neurons that respond to both A-∂ and C neurons, as well as to other peripheral stimuli. WDR neurons are also sensitive to visceral stimuli. It is possible that referred and sympathetic pains depend upon this sort of visceral-somatic convergence.[17]

---

[15] "Sensory Disturbances from Cerebral Lesions," *Brain*, xxxiv (1911): 102-254.

[16] Information traveling at the slower speeds would take about 8 seconds to reach a horse's spinal column from its hoof.

Once pain information exits the dorsal horn, it travels either to the reticular formation in the brain stem or to the thalamus. Laminae I and V project to the lateral nuclei in the thalamus,[18] and laminae I, V, and VI project to the medial nuclei. Each type of nuclei underwrites a different sort of information; the lateral nuclei process discriminative information (fast pain), while the medial nuclei and reticular connections process affective-motivational information (slow pain). The two thalamic streams remain separate on their trip to the cortex as well. Pain neurons in the lateral nuclei synapse in the somatosensory cortex, which then can compute the location and characteristics of the pain; those in the medial nuclei synapse in the anterior cingulate gyrus in the frontal lobe, which figures in our emotional reactions to pain. The frontal lobe (and its connections) process our actual suffering.

III. PHILOSOPHY'S ERROR

Now we can see how and why several philosophers are mistaken in their conclusions that there are no such things as pains,[19] or that pains are located in our limbs,[20] or that pains are purely subjective,[21] or that pains are reactive behaviors.[22] Each of these positions identifies pain with one of the neuronal groups within the pain system, while failing to recognize that our pain system is complex and contains at least a duality[23] of subsystems, each of which processes a different sort of information. (See figure 2.) In general, philosophers make these mistakes because they misunderstand the double dissociation methodology. We can, either through purposeful intervention or accidents of nature, dissociate our discriminative pain processing from our affective-motivational pain processing. Ingestion of morphine (or other opiates), lesions to the medial thalamus, and prefrontal lobotomies all result in sensations of pain without a sense of suffering and without producing characteristic pain behaviors (wincing, moaning, complaining, and so on) (*ibid.*). In these cases, patients can localize their pains but are not upset by the

---

[17] For contrary evidence, though, see G.D. Schott, "Visceral Afferents: Their Contribution to 'Sympathetic Dependent' Pain," *Brain*, CXVII (1994): 397-413. Other neurons that show this sort of convergence are the "nociceptive-specific" neurons in lamina I and the "complex" neurons in laminae VII and VIII of the dorsal horn. The interaction of autonomic information with somatic appears to be quite common throughout our pain system.

[18] A.D. Craig, M.C. Bushnell, E-T. Zhang, and A. Blomqvist, "A Thalamic Nucleus Specific for Pain and Temperature Sensation," *Nature*, CCCLXXII (1994): 770-73.

[19] P.M. Churchland, "Reduction, Qualia, and the Direct Introspection of Brain States," this JOURNAL, XXXVIII, 1 (January 1985): 8-28; Dennett, "Why You Can't Make a Computer That Feels Pain," *Synthese*, XXXVIII (1978): 449.

[20] D.M. Armstrong, *The Nature of Mind and Other Essays* (Ithaca: Cornell, 1981); N. Newton, "On Viewing Pain as a Secondary Quality," *Noûs*, XXIII (1989): 569-98; G. Pritcher, "Pain Perception," *Philosophical Review*, LXXIX (1970): 368-93.

[21] G.R. Gillett, "The Neurophilosophy of Pain," *Philosophy*, LXVI (1991): 191-206; N. Grabek, "Objective and Subjective Aspects of Pain," *Philosophical Psychology*, IV (1991): 249-66; C. McGinn, *The Subjective View* (New York: Oxford, 1983).

[22] L. Wittgenstein, *Philosophical Investigations*, G.E.M. Anscombe, trans. (Cambridge: Blackwell, 1953).

[23] R. Melzack and P.D. Wall, "Pain Mechanisms: A New Theory," *Science*, CL (1965): 971-79, and E.R. Hilgard and J.R. Hilgard, *Hypnosis in the Relief of Pain* (New York: Brunner/Mazel, 1994, rev. ed.) both argue that there are three components to pain processing: the discriminative, the affective-emotional, and the evaluative. I find little physiological evidence to support these claims. Here, I discuss only two components.

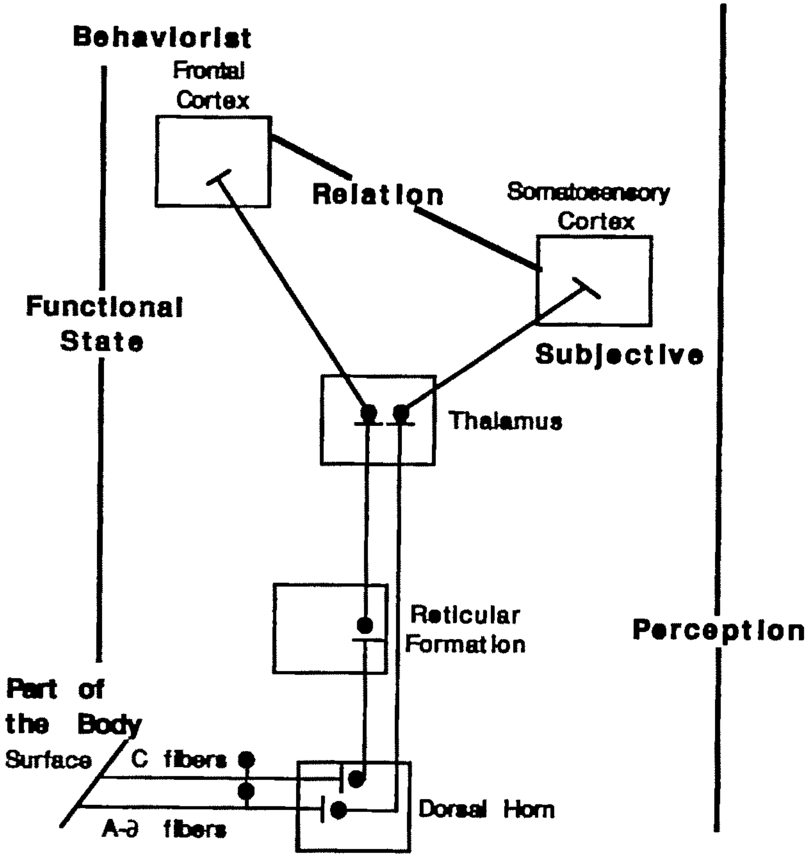


Figure 2. Part of the pain system identified with different philosophical views regarding the basic nature of pain. Most views identify pain with one subsystem or with one sort of neural processor. Two exceptions are functional state and perceptual views of pain, though perceptual views of pain often overlook or underestimate the motor component of pain processing.

fact that they are in pain. We can also get reverse effects, to a degree. Fentanyl causes one to react in pain, yet inhibits our discriminatory abilities for the pain.[24] Lesion studies and studies using hemispherectomies show that even with the cortex completely missing, we can still have a pain sensation; we simply lack fine localization and intensity discrimination.[25] Patients with Parkinson's disease and Huntington's chorea often have pain sensations but are unable to indicate where they feel the pains.[26]

We also find instances of the pain centers in the thalamus and cortex being activated without corresponding activations of A-∂ or C fibers ("nociception"). Fully 80% of lower back pain sufferers present no external or internal injury.[27] Phantom limbs and phantom pains in phantom limbs are quite common experiences in new amputees.[28] Stimulating the medial periaqueductal gray region, tectum, or thalamus directly can also result in painful experiences.[29] Finally, our emotional states heavily influence the degree of pain we feel, quite independent of actual injury. Indeed, psychogenic pains, pains without any corresponding injury or peripheral stimulation, have been documented for quite some time.[30]

---

[24] R.H. Gracely, R. Dubner, and P.A. McGrath, "Fentanyl Reduces the Intensity of Painful Tooth Pulp Sensations: Controlling for Detection of Active Drugs," *Anesthesia and Analgesia*, LXI (1982): 751-55.

[25] K.D. Davis, R.R. Tasker, Z.H.T. Kiss, W.D. Hutchison, and J.O. Dostrovsky, "Visceral Pain Evoked by Thalamic Microstimulation in Humans," *Neuroreport*, VI (1995): 369-74; L.A. French, S.N. Chou, and J.L. Story, *Clinical Orthopedics*, XLVI (1966): 83-86; Head and Holmes; and Roland.

[26] E.H. Chudler and W.K. Dong, "The Role of the Basal Ganglia in Nociception and Pain," *Pain*, LX (1995): 3-85.

[27] Wall, "Introduction," in Wall and Melzack, eds., *Textbook of Pain* (New York: Churchill Livingstone, 1989, 2nd ed.), pp. 1-18.

[28] See discussion in Hilgard and Hilgard; T.S. Jensen and P. Rasmussen, "Phantom Pain and Related Phenomena After Amputation," in Wall and Melzack, eds., pp. 508-21.

[29] Davis et alia; K.A. Keay, C.I. Clement, B. Owler, A. Depaulis, and R. Bandelr, "Convergence of Deep Somatic and Visceral Nociceptive-Information onto a Discrete Ventrolateral Midbrain Periqueductal Gray Region," *Neuroscience*, LXI (1994): 727-32.

[30] R. Roy, "Engel's Pain-Prone Disorder Patient: 25 Years After," *Psychotheraphy Psychosomatics*, XLIII (1985): 126-35. In *Illustrations of the Influence of the Mind upon the Body in Health and Disease Designed to Elucidate the Imagination* (Philadelphia: Henry C. Lee's, 1884, 2nd ed.), D.H. Tuke reported the case of a butcher who got fouled up on a meat hook and appeared to be in agony. When examined by the local chemist, however, it was discovered that the meathook had only penetrated his jacket sleeve and, even though the butcher was screaming in "excessive pain," he was completely unharmed (p. 168, as reported in A. Gamsa, "The Role of Psychological Factors in Chronic Pain, I: A Half Century of Study," *Pain*, LVII (1994): 5-15).

Correlatively, there are also examples of our peripheral pain processors being activated without this information proceeding on to the thalamus or cortex. About 37% of emergency-room visitors felt no pain at the time of their injury.[31] Athletes and soldiers can continue performing free of pain, even though they have been severely injured.[32] Hypnosis allows some subjects to engage in what would otherwise be painful activities without being in pain.[33] Placebos are notoriously helpful in relieving pain. (Interestingly enough, they relieve pain at half the rate of the real drug, regardless of the supposed strength of the drug.[34]) Of course, some lesions to the thalamus and cortex can result in the cessation of pain experiences, even though the peripheral neurons continue to operate normally.

Each of these double dissociations, however, individuates neuronal groups or subsystems only *within* our overarching pain system. I make this claim by analogy with our other perceptual systems. Blindsight patients can discriminate shapes and figures they claim not to be able to see consciously. Sufferers of Anton's syndrome insist that they can see perfectly well, even though they are completely blind and have severe bilateral damage to their visual association areas. No one uses these facts to argue, however, that vision does not exist, or that vision is located in our eyeballs, or that vision is purely subjective, or that vision is behavioral. We may not know exactly what to say about blindsight or Anton's syndrome, but no one claims that blindsight is not a disorder of the visual system or that patients with Anton's syndrome are not having a visual experience of some sort.

By misunderstanding what a perceptual system in the brain encompasses, many philosophers miss the boat regarding the basic nature and structure of pain. Double dissociation alone does not individuate our basic systems; that is used to isolate the subsystems that operate within the larger system. We then need to build our different systems out of the component pieces. Teleological considerations help us to do so. To wit: the neurons in our pain system all respond to roughly the same sort of information; they increase their rate of firing in the presence of noxious stimuli on skin or deep organs. Moreover, the connection among the six-neuron tract is a stable, common, and isolable pathway. Connections fore and aft show a stream of information flowing from the nociceptors on the skin up through the cortex.[35] Finally, a pain sensory system tied to the somatosensory processors makes good evolutionary sense. As creatures eking out lives in a hostile environment, having a system which could warn us when damage occurred and which could force us to protect damaged parts until they healed would be tremendously beneficial. (Indeed, persons who cannot feel any pain at all often live a nasty, brutish, and short life.[36])

---

[31] Melzack, Wall, and T.C. Ty, "Acute Pain in an Emergency Clinic: Latency of Onset and Descriptor Patterns Related to Different Injuries," *Pain*, XIV (1982): 33-43.

[32] H.K. Beecher, "Relationship of Significance of Wound to the Pain Experience," *Journal of the American Medical Association*, CLXI (1956): 1609-13.

[33] Evoked potential recordings of painful stimuli under hypnosis indicates that at least activity in the frontal lobe is affected (Helen Crawford, personal conversation).

[34] F.J. Evans, "The Placebo Response in Pain Reduction," in J.J. Bonica, ed., *International Symposium on Pain, Advances in Neurology, Volume IV* (New York: Raven, 1974), pp. 289-96.

Neither our conscious experience of pain, the damaged tissue itself, nor our bodily or emotional reactions are fundamental to pain processing. Each is but one component of a larger processor. Hence, it is a mistake to try and claim one or the other as pain *simpliciter*. It is equally erroneous to conclude that since we cannot identify one or the other with pain, there is no such thing. The entire pain sensory system functions largely the same as any of our sensory systems. Their pieces are united by our best guess of their function, based on the three types of converging evidence discussed above. Hence, we have concluded that the components of our visual system take the information contained in photons bouncing around in the world and use it to compute the location, orientation, texture, color, and movement of objects in the environment. The components of our auditory system take the information contained in atmospheric compression waves and use it to compute the placement of things. And the components of our pain system take pressure, temperature, and chemical readings of our surface (and interior) and use this information to track what is happening to our tissues. The A-∂ cells and the C fibers do this, as do the spinothalamic tract and its connections to the cortex. In sum, we have a complex but well-defined sensory system that monitors our tissues to promote the welfare of our bodies.[37]

[35] Exactly what processing algorithms are being executed is a more complicated story, and one heavily influenced by the pain inhibitory system. Accounting for the details of the computations is beyond the scope of this paper.

[36] Critchley, "Congenital Indifference to Pain," *Annals of Internal Medicine*, XLV (1956): 737-47; R.A. Sternbach, "Congenital Insensitivity to Pain: A Critique," *Psychological Bulletin*, LX (1963): 252-64; and *Pain: A Psychophysiological Analysis* (New York: Academic, 1968).

[37] See also Wall. This view, too, accounts for why our C-fiber systems might be so slow. General monitoring of bodily conditions should not often require a quick response.

IV. STRANGE FACTS ABOUT PAIN

As I stated in the introduction, the above is an easy and less interesting reason why philosophers are wrong. More to the point: this way of looking at our pain system does not clear up all of the confusions, mysteries, and conflicts in our account of pain. There are several important empirical facts that any theory of pain needs to be able to explain. These are facts that, by and large, do not have analogues in our other perceptual systems. Moreover, they are facts that lead otherwise intelligent people to make prima facie bizarre statements regarding what pain is and is not. Here, I can only touch upon a few such facts, but it should be enough to motivate the challenge to theories of pain as well as to justify the approach I shall advocate in understanding pain phenomena toward the end of this paper.

(1) There is, in fact, a poor correlation between nociception and pain perception.[38] That is, the relationship between stimulating the A-∂ and C fibers and actually feeling or reporting a pain is not at all straightforward. Several tribal rituals give vivid illustrations of the dissociation. In parts of India, for example, men chosen to represent the gods have steel hooks inserted under the muscles of their back. They then swing above the crowds, suspended on these hooks by ropes, blessing children and crops. They exhibit no pain.[39] I mentioned that about 40% of all emergency-room patients reported feeling no pain at the time of injury; 40% more report greater pain than one would expect, leaving only 20% of all emergency-room visitors having pains appropriate to their injuries.[40] It is not the case that we can dissociate nociception from discriminative and affective-motivational reaction; it is that they regularly and frequently dissociate.

Our other perceptual systems are not like this. There is a highly reliable correlation between having the rods and cones in our retina being bombarded by light photons and our having some visual experience or other.[41] There is also a highly reliable correlation between the vibration of our tympanic membranes at a certain frequency and hearing sounds. Any theory of pain is going to have to explain why our peripheral sensors for noxious stimuli do not appear to be well-connected to our sensations of pain. Indeed, relation between external events and internal indicators is part of what individuates our systems. Without better correlation between an external event and internal activity, we cannot claim that our putative pain system is, in fact, a pain system.

---

[38] Wall, "The Dorsal Horn," in Wall and Melzack, eds., pp. 102-11; Wall and S.B. McMahon, "Microneuronography and Its Relation to Perceived Sensation," *Pain*, XXII (1985): 209-29.

[39] D.D. Kosambi, "Living Prehistory in India," *Scientific American*, CXVI (1967): 106-14.

[40] See Melzack et alia.

[41] It is a live debate whether there is a good correspondence between the pattern of activity in the retina and a particular *type* of experience. How you answer that question depends upon how elastic you believe our visual "module" to be. P.M. Churchland and R.L. Gregory, for example, think that our visual system is cognitively penetrable from above; J. Fodor does not.

(2) Chronic pains are, by and large, a mystery.[42] Some people have pains which last for years, with no discernible cause, and which are completely resistant to treatment. These are cases above and beyond things like phantom pains (which also have no discernible cause and are resistant to treatment), for one might argue that phantom pains are simply abnormal instances of human suffering. (Who knows what is supposed to happen if you lose something as large as a limb? Surely, many areas of the topographically arranged somatosensory system will be thrown off track; maybe pains that appear to be in the limb that no longer exists are not so strange after all.[43]) But with chronic pains, otherwise perfectly normal people—with no serious (or even superficial) injury—live their lives in constant pain. Moreover, removing bits of the spinal column, the dorsal horn, the thalamus, the reticular formation, the somatosensory cortex, or the frontal lobe concerned with pain have no effect on the patients being in pain.

Again, there are few parallels with other perceptual systems. Rarely do otherwise normal individuals have ongoing visual or auditory experiences without some determinate cause and explanation. Chronic hallucinations by themselves are quite rare.

(3) Low-level stimulation of our thermoreceptors (the larger A-β fibers), which are not supposed to be connected to pain perception, inhibits the experience of pain,[44] while a higher level stimulus exac-

---

[42] K.L. Casey, S. Minoshima, R.A. Koeppe, J. Weeder, and T.J. Morrow, "Temporo-Spatial Dynamics of Human Forebrain Activity During Noxious Heat Stimulation," *Society of Neuroscience Abstracts*, xx (1994): 1573; C.A. Pagni, "Central Pain Due to Spinal Cord and Brian Stem Damage," in Wall and Melzack, eds., pp. 634-55.

[43] Some phantom pains may be attributable to memories of recent pains in the amputated limb, for phantom pains bear an uncanny resemblance to pre-amputation experiences of pain. See J. Katz, "Psychophysiological Contributions to Phantom Limbs," *Canadian Journal of Psychiatry*, xxxvii (1992): 282-98; E.J. Krane and L.B. Heller, "The Prevalence of Phantom Sensation and Pain in Pediatric Amputees," *Journal of Pain Symptom Management*, x (1995): 21-29; Melzack, "Phantom Limbs, the Self, and the Brain," *Canadian Psychology*, xxx (1989): 1-16.

[44] Wall, "Presynaptic Control of Impulses at the First Central Synapse in the Cutaneous Pathway," *Physiology of Spinal Neurons: Progress in Brain Research XII* (New

erbates pain.[illegible] Here, there are some parallels with our other systems, if we take the A-β interference to be a type of hard-wired pain illusion. Similarly, we see a straight stick as bent when placed half way in water, and we arrange groups of dots into rows and columns.

Facts such as these led the International Association for the Study of Pain (IASP) Subcommittee on Classification to conclude: "Pain is always subjective.... Many people report pain in the absence of tissue damage or any pathophysiological cause; usually this happens for psychological reasons. There is usually no way to distinguish their experience from that due to tissue damage if we take the subjective report... [P]ain...is always a psychological state)."[illegible] Could they be correct? Are the connections between actual tissue damage, or some other injury, and our sensation of pain so weak that it is better to discount nociception entirely when defining pain? The IASP subcommittee clearly thinks so in their definition of pain: "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, *or described in terms of such damage*" (*ibid.*; italics added).

This sort of position is not preferable, for a variety of reasons. First, if pains are not correlated with actual injury, or the potential for damage, then we lose our intuitive evolutionary story about why we have a pain-sensing system. If our pain system has somehow become detached from the job it is supposed to perform, then its existence and poor performance no longer have clear explanations.

Second, pains become very peculiar phenomena indeed, quite unlike our other qualitative experiences. We can have visual or auditory hallucinations—we can be mistaken about what we think we are perceiving—but if pain is purely subjective, then there is no way for us to have an illusion of being in pain. Phantom pains become just regular pains instead of some special case demanding special consideration and treatment. Of course, saying something is strange is not a reason for saying that it does not exist, but it is a reason, I believe, to be cautious in making such metaphysical commitments. Indeed, I do

---

York: Elsevier, 1964), pp. 92-118; Wall and Cronly-Dillon, "Pain, Itch, and Vibration," *Archives of Neurology*, II (1960): 365-75; Wall and W.H. Sweet, "Temporary Abolition of Pain in Man," *Science*, CLV (1967): 108-09; see also W.D. Willis and R.E. Coggeshall, *Sensory Mechanisms in the Spinal Cord* (New York: Plenum, 1978); T.L. Yaksh, *Spinal Afferent Processing* (New York: Plenum, 1986).

[illegible] J.C. Willer, F. Boureau, and D. Albe-Fessard, "Human Nociceptive Reactions: Effects of Spatial Summation of Afferent Input from Relatively Large Diameter Fibers," *Brain Research*, CCI (1980): 465-70.

[illegible] International Association for the Study of Pain (IASP) Subcommittee on Classification, "Pain Terms: A Current List with Definitions and Notes on Usage," *Pain*, Supplement 3 (1986), p. 217.

think that the IASP subcommittee (and Colin McGinn and Saul Kripke) are quite wrong in their understanding of what pain is.

V. GATE THEORIES OF PAIN

In general, gate theories of pain argue that the robust feedback loops in our pain system serve to inhibit, enhance, or distort incoming nociceptive information. So, in addition to the basic six-neuron structure of our pain system I outlined above, there are additional ascending connections from the spinal cord to the brain stem, circular pathways from the spinal cord to other areas in the spinal cord itself, and descending feedback loops from the cortex, hypothalamus, and brain stem back to the spinal cord.[47] Some portions of this theory have been worked out in considerable detail. Some of the mysterious pain phenomena can be explained in terms of well-confirmed portions of Melzack and Wall's[48] gate theory of pain.

Take the example of altering our sensation of pain by stimulating the larger A-β fibers. The A-∂ and C neurons in the dorsal horn are connected via inhibitory interneurons. These interneurons are stimulated by low-threshold A-β fibers. This means that stimulating A-β cells dampens the activity of the A-∂ and C neurons in the dorsal horn, which in turn means that less pain information would travel up to the brain.[49] (See figure 3.) Or, another example: gently pressing a grid with alternating cool and warm bars on the skin often causes the sensation of a strong burning pain.[50] Neurophysiological recordings now show that there are several ascending neurons that are sensitive to both pain and temperature. A bit of central disinhibition plus these bimodal neurons explain how this illusion can occur. It can also explain why cold things burn.[51] The relationship between tactile stimulation and pain is not completely clear here, however. Recent evidence indicates that painful stimuli can substantially decrease our sensations of touch.[52]

---

[47] See also Apkarian.

[48] Wall, "Introduction"; Melzack and Wall, "Pain Mechanisms," and *The Challenge of Pain* (New York: Penguin, 1986); see also N.F. Britton and S.M. Skevington, "A Mathematical Model of the Gate Control Theory of Pain," *Journal of Theoretical Biology*, CXXXVII (1980): 91-105. Several aspects of the original theory have been shown to be incorrect; for discussion, see P.W. Nathan, "The Gate-Control Theory: A Critical Review," *Brain*, XCIX (1976): 123-58; and Kandel and Schwartz. Most accept the general outline of the view, however.

[49] Wall, "Introduction."

[50] Known as Thunberg's thermal grill illusion, this was first demonstrated in 1896.

[51] A.D. Craig and M.C. Bushnell, "The Thermal Grill Illusion: Unmasking the Burn of Cold Pain," *Science*, CCLXV (1994): 252-55.

[52] Apkarian, R.A. Stea, and S.J. Bolanowski, "Heat-Induced Pain Diminishes Vibrotactile Perception, A Touch Gate," *Somatosensory Motor Research*, XI (1994): 259-67.

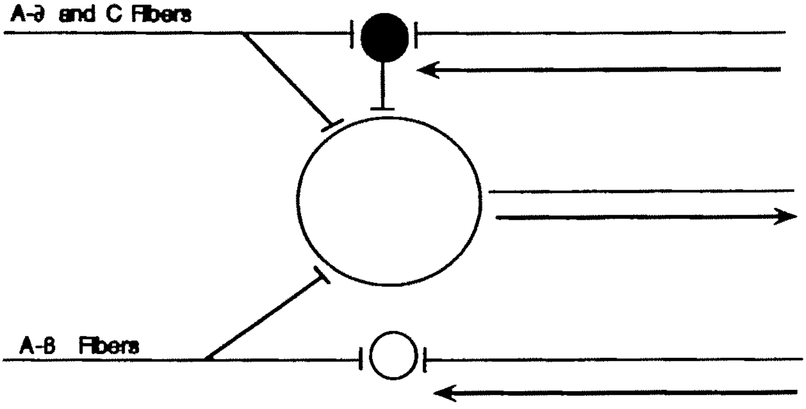


Figure 3. Low-level gate control. Cells in the dorsal horn and trigeminal nuclei which respond to A-∂- and C-fiber input have excitatory (filled circle) and inhibitory (open circle) interneurons associated with them. A-β fibers activate the inhibitory neurons as well as input to the nociceptive cells. Descending controls input to the interneurons.

Moreover, though more research needs to be done, the alternate routes in the central nervous system might help explain why cutting the primary dorsal horn pain channel, or parts of the spinal cord itself, may not remove chronic pain.[33] Indeed, it looks as though the failure at curing chronic pains by lesions is overstated anyway.[34] Some neuropathic pains are controllable using subarachnoid opioids, as they interfere with the pain centers in the dorsal horn, which are probably responsible for deafferantation pain.[35] Other data indicate that most chronic or prolonged pains are the result of changes in central nervous functions that have been caused by the neural firing patterns of early nociception.[36]

A Melzack-Wall type of gate-control theory of pain cannot be the entire story, however, for three reasons. First, we do not have anything approaching a complete theory at this stage in the game. Although the theory does explain low-level phenomena, nothing has been worked out in particular regarding the clearly "psychological" influences.[87] Melzack and Wall themselves simply gesture toward a central gating mechanism that presumably would explain hypnosis effects, any remaining chronic pains, the dismal correlations between emergency-room injuries or back injuries and pain, and so on. But without providing more details—where this central gating mechanism is located in the brain, how it functions to alter our pain perceptions, why we have such a powerful mechanism, for example—then they have said little more than that "some other stuff happens in the head that explains everything else." As Wall himself remarks of the many areas implicated in inhibiting our sensations of pain: "unfortunately we know little of their relative importance and nothing of the actual circumstances in which they come into action."[88]

---

[33] S. Canavero, "Dynamic Reverberation: A Unified Mechanism for Central and Phantom Pain," *Medical Hypotheses*, XVII (1994): 203-07; see also A. Dray, L. Urban, and A. Dickenson, "Pharmacology of Chronic Pain," *Trends in Pharmacological Sciences*, XV (1994): 190-97.

[34] See Canavero.

[35] R.P. Iacono, M.V. Boswell, and M. Neumann, "Deafferentation Pain Exacerbated by Subarachnoid Lidocaine and Relieved by Subarachnoid Morphine: Case Report," *Regional Anesthesiology (United States)*, XIX (1994): 212-15.

[36] See discussion and review in A.L. Vaccarino and D.A. Chorney, "Descending Modulation of Central Neural Plasticity in the Formalin Pain Test," *Brain Research*, DCLXVI (1994): 104-08.

Second, and more importantly, even if we could get the details on some sort of central gating mechanism, this would not mean that pains are not largely subjective (which is but a step away from being purely subjective). If top-down cortical processes (which is what I take 'purely subjective' to mean here) are mainly responsible for our sensations of pain, then the IASP subcommittee would be right and pains would be deeply peculiar. They would be the only perceptual experiences we have which are normal (that is, normal functioning creatures have them), natural (that is, not the product of tweaking something in a laboratory, but occurring in the wild, as it were), commonplace (we have them all the time), believed to be giving us information about the external environment (external to the brain, that is (but internal to our bodies)), but in fact not. Dreams might be the only exception, but, unlike dreams, we do not realize when a pain is over that it really was not about the external world after all.

Finally, "the body of psychological research into pain has failed to yield compelling evidence for a direct causal relationship between psychological factors and pain in the general population of pain patients."[89] In general, if physicians cannot find a physical cause for a pain and it continues despite medical interventions, then it is attributed to something "psychological" (*ibid.*).[20] This can be at best an unsupported hypothesis, however. At worst, it is a euphemism for our own ignorance. If pains are driven by psychological factors, then we should be able to isolate what those factors are. Thus far, we have been unsuccessful in discovering any psychological trait or ailment that is correlated with sensations of pain. In sum: those who claim that pain is subjective or psychological are being too facile.

---

[87] The influence of the psychological on the perception of pain has a long and venerable history; see Whytt (1786), Brodie (1837), Carter (1853), as described in H. Merskey and F.G. Spear, *Pain: Psychological and Psychiatric Aspects* (London: Baillière, Tindall, and Cassell, 1967).

[88] Wall, "Introduction," p. 12.

[89] A. Gamsa, "The Role of Psychological Factors in Chronic Pain, II: A Critical Appraisal," *Pain*, LVII (1994): 23.

Most philosophers, psychologists, and neurophysiologists who do not fall prey to the less interesting mistake do hold some version of a gate-control theory. Some advocate it explicitly (for example, Dennett, Melzack, Wall); others maintain it implicitly (Hilgard, Tye). Either way, I believe this commitment to be premature. If there is any way for us not to hold that our pain system is purely, or largely, or even significantly subjective, then we should not do it. I propose that instead of acquiescing to the IASP subcommittee's conclusion, we should take the mysterious and unexplained pain phenomena as evidence that the six-neuron view (even with the additional bells and whistles) is wrong. I suggest that we should reexamine the data from a different perspective.

VI. A PAIN-INHIBITING SYSTEM

If standard (higher level) gate-theory approaches are misguided, then we still have a chance at understanding pain from a solid neurophysiological and biological perspective without making pain into something peculiar. As will be clear in a moment, I advocate dividing what Melzack et alia have lumped into one cortically-driven pain system into two separate systems: a nociceptor-driven pain sensory system (PSS) and a largely top-down pain inhibitory system (PIS). First, though, let me sketch some data that I believe are relevant to this perspective, data that the more traditional gate theories either overlook or minimize; I shall then outline some theoretical considerations that point toward a two-system model.

Proponents of gate theories write as though just about any psychological event or any area of the cortex has the potential of influencing the perception of pain, what Melzack calls a *neuromatrix*.[61] Although there are lots of feedback loops and other sorts of pain connections, not every area in the brain is sensitive to pain information. Recent advances in functional imaging technology—which includes magnetic resonance imaging (MRI), xenon gamma emission detection, single photon computer emission tomography (SPECT), and positron emission tomography (PET)—allow us to map active areas of the brain with greater precision than ever before.[62] Imaging studies of pains clearly show that chronic pains are correlated with increased activity in the cingulate and frontal cortex, as well as sometimes with the insular cortex, hypothalamus, and periaqueductal gray.[63] Phasic pains are keyed to increased activity in the anterior cingulate, frontal cortex, thalamus, and primary and secondary somatosensory cortex.[64] Although lack of activity does not prove conclusively that an area is not sensitive to pain information, these sorts of studies should give advocates of global cortical influences on our perception of pain pause for thought.

---

[20] R.C. Kupers, H. Konings, H. Andriaensen, and J. Gybels, "Morphine Differentially Affects the Sensory and Affective Pain Rating in Neurogenic and Idiopathic Forms of Pain," *Pain*, XLVII (1991): 5-12; see also discussion in S. Benjamin, D. Barnes, S. Berger, I. Clarke, and J. Jeacock, "The Relationship of Chronic Pain, Mental Illness, and Organic Disorders," *Pain*, XXXII (1988): 185-91; M. Grushka, B.J. Sessle, and R. Miller, "Pain and Personality Profiles in Burning Mouth Syndrome," *Pain*, XXVIII (1987): 155-67; H. Merskey, "Symptoms that Depress the Doctor, Too Much Pain," *British Journal of Hospital Medicine* (January 1984): 63-66, and "Psychiatry and Chronic Pain," *Canadian Journal of Psychiatry*, XXXIV (1989): 329-35; R.A. Sherman, C.J. Sherman, and G.M. Bruno, "Psychological Factors Influencing Chronic Phantom Limb Pain: An Analysis of the Literature," *Pain*, XXVIII (1987): 285-95.

---

[61] Melzack, "Phantom Limbs and the Concept of a Neuromatrix," *Trends in Neuroscience*, xiii (1990): 88-92; "Central Pain Syndromes and Theories of Pain," in E.L. Casey, ed., *Pain and Central Nervous Disease: The Central Pain Syndromes* (New York: Raven, 1991), pp. 59-64, and "Phantom Limbs," *Scientific American*, cclxvi (1992): 90-96; see also discussion in Canavero.

[62] See Apkarian.

[63] Y. Hosobuchi, "Treatment of Cerebral Ischemia with Electrical Stimulation of the Cervical Spinal Cord," *Pace*, xiv (1991): 122-26; Y. Katayama, T. Tsubokawa, T. Hirayama, G. Kido, T. Tsukiyama, and M. Iio, "Response of Regional Cerebral Blood Flow and Oxygen Metabolism to Thalamic Stimulation in Humans as Revealed by Positron Emission Tomography," *Journal of Cerebral Blood Flow Metabolism*, vi (1986): 637-41; E.C. La Terre, A.G. De Volder, and A.M. Goffinet, "Brain Glucose Metabolism in Thalamic Syndrome," *Journal of Neurosurgery and Psychiatry*, li (1988): 427-28; Y.R. Tran Dinh, C. Thurel, A. Serrie, G. Cunin, and J. Seylaz, "Glycerol Injection into the Trigeminal Ganglion Provokes a Selective Increase in Human Cerebral Blood Flow," *Pain*, xlvi (1991): 13-16.

[64] Apkarian; Apkarian, R.A. Stea, S.H. Manglos, N.M. Szeverenyi, R.B. King, and F.D. Thomas, "Persistent Pain Inhibits Contralateral Somatosensory Cortical Activity in Humans," *Neuroscience Letters*, cxl (1992): 141-47; M. Backonja, E.W. Howland, J. Wang, J. Smith, M. Salinsky, and C.S. Cleeland, "Tonic Changes in Alpha Power During Immersion of the Hand in Cold Water," *Electroencephalography and Clinical Neurophysiology*, lxxix (1991): 192-203; E.L. Casey, ed. *Pain and Central Nervous Disease*; K.L. Casey, S. Minoshima, K.L. Berger, R.A. Koeppe, T.J. Morrow, and K.A. Frey, "Positron Emission Tomographic Analysis of Cerebral Structures Activated Specifically by Repetitive Noxious Heat Stimuli," *Journal of Neurophysiology*, lxxi (1994): 802-07; R.C. Coghill, J.D. Talbot, A.C. Evans, E. Meyer, A. Gjedde, M.C. Bushnell, and G.H. Duncan, "Distributed Processing of Pain and Vibration by the Human Brain," *Journal of Neuroscience*, xiv (1994): 4095-108; V. Di Piero, S. Ferracuti, U. Sabatini, P. Pantano, G. Cruccu, and G.L. Lenzi, "A Cerebral Blood Flow Study on Tonic Pain Activation in Man," *Pain*, lvi (1994): 167-73; G.H. Duncan, C. Morin, R.C. Coghill, A. Evans, K.J. Worsley, and M.C. Bushell, "Using Psychophysical Ratings to Map the Human Brain Regression of Regional Cerebral Blood Flow (RCBF) to Tonic Pain Perception," *Society of Neuroscience Abstracts*, xx (1994): 1572; J. Hsie, Ö. Hägermark, M. Ståhle-Bäckdahl, K. Ericson, L. Eriksson, S. Stone-Elander, and M. Ingvar, "Urge to Scratch Represented in the Human Cerebral Cortex During Itch," *Journal of Neurophysiology*, lxxii (1994): 3004-08; A.K.P. Jones,

Moreover, the neuronal areas sensitive to pain information are different from what we originally believed. Rat studies suggest that, in addition to the structures discussed above, areas of the limbic system are also involved.[65] Functional images of human brains indicate that homologous areas are involved in us as well.[66] Moreover, limbic activity in chronic pains becomes quite important as imaging studies also show a *decrease* in thalamic activity, instead of the increase one would expect. The decrease in thalamic activity, coupled with the increase in cingulate response, tells us that a spinothalamic pathway cannot be what is causing the increase. Scientists have discovered a direct connection between the spinal-cord projections and multiple limbic areas in both the rat and monkey.[67] The limbic system is then tightly connected with the cingulate and frontal cortex.

The limbic system also receives inputs from the reticular system (another pain juncture). As Apkarian concludes, "the brain imagining studies of pain...point to a very different emphasis in research regarding the central processing of pain.... [S]ystems outside of the spinothalamic system may control the type of processing taking place in the spinothalamic system. [Chronic pains]...seem to activate cortical areas outside the spinothalamic domain, which in turn inhibit the spinothalamic inputs to the cortex" (*op. cit.*, p. 290). That the thalamus and cortex are probably not important in processing our perceptions of pain stands in direct contrast to what many take to be common knowledge regarding how our pain system works.

Because new facts regarding the transmission of pain information in the brain are only slowly emerging, we should be fairly conservative in our conclusions. All I want to claim at the moment is that the

---

W.D. Brown, K.J. Friston, L.Y. Qi, and R.S.J. Frackowiak, "Cortical and Subcortical Localization of Response to Pain in Man Using Positron Emission Tomography," *Proceedings of the Royal Society of London*, CCXLIV (1991): 39-44; J.D. Talbot, S. Marrett, A.C. Evans, E. Meyer, M.C. Bushnell, and G.H. Duncan, "Multiple Representations of Pain in Human Cerebral Cortex," *Science*, CCLI (1991): 1355-58.

[65] J. Mao, D.J. Mayer, and D.D. Price, "Patterns of Increased Brain Activity Indicative of Pain in a Rat Model of Peripheral Mononeuropathy," *Journal of Neuroscience*, XIII (1993): 2689-702.

[66] See Apkarian; and Roland.

[67] Apkarian; H.M. Newman, R.T. Stevens, C.M. Power, A.V. Apkarian, "Spinal-Suprathalamic Projections from the Upper Cervical and the Cervical Enlargement in Rat and Squirrel Monkey," *Society of Neuroscience Abstracts*, XX (1994): 118.

six-neuron sketch of nociception transmission is probably over simplified to the point of being misleading. Accepting this conclusion opens the possibility of describing our pain systems differently than it has been in the past. More important for our purposes, it allows the possibility of dividing what we have been calling our pain system into two different and independent processing streams. I shall argue that doing so helps clear up the remaining puzzles in pain phenomena.

We have known for some time that many of the inhibitory streams are not merely feedback loops in our ascending pain fibers, for they are anatomically distinct from our pain processors. Three areas are primarily responsible for inhibiting pain information in the spinal column: the cortex, the thalamus, and the brain stem. The dorsal raphe is probably heavily involved as well.[68] In particular, the neocortex and hypothalamus project to the periaquaductal gray region (PAG), which then sends projections to the reticular formation (see figure 4). The reticular nuclei then work to inhibit activity in the dorsal horn.[69] This processing stream works by preventing a central cortical representation of pain from forming.[70] Endogenous opioids, stimulating the PAG, and morphine all dampen incoming information in the same way in the dorsal horn. That is, this pain inhibition stream does not merely disrupt the transmission of pain information, it actively prevents it from occurring.

In addition, this subsystem is not the only pain inhibitor at work. Stress-induced analgesia can occur without any opioids being released and it is not prevented from occurring when opioid-blockers are administered.[71] Moreover, different neural substrates are involved in inhibiting fast pains and slow pains. Stimulating the hypothalamus reduces tonic pain and is not related to stress-related analgesia.[72]

---

[68] Q.-P. Wang and Y. Nakai, "The Dorsal Raphe: An Important Nucleus in Pain Modulation," *Brain Research Bulletin*, VI (1994): 575.

[69] H.L. Fields, "An Endorphin-Mediated Analgesia System, Experimental and Clinical Observations," in J.B. Martin, S. Reichlin, and K.L. Brick, eds., *Neurosecretion and Brain Peptides, Implications for Brian Function and Neurological Disease* (New York: Raven, 1981); H.L. Fields and A.I. Basbaum, "Endogenous Pain Control Mechanisms," in Wall and Melzack, eds., pp. 206-19; for references and reivew, see also J.E. Sherman and J.C. Liebeskind, "An Endorphinergic, Centrifugal Substrate of Pain Modulation: Recent Findings, Current Concepts, and Complexities," in J.J. Bonica, ed., *Pain* (New York: Raven, 1980), pp. 191-204.

[70] See Vaccarino and Chorney.

[71] See Kandel and Schwartz; and Fields and Busbaum.

[72] R. Lopez, S.L. Young, and V.C. Cox, "Analgesia for Formalin-Induced Pain by Lateral Hypothalamic Stimulation," *Brain Research*, DLXIII (1991): 1-7.

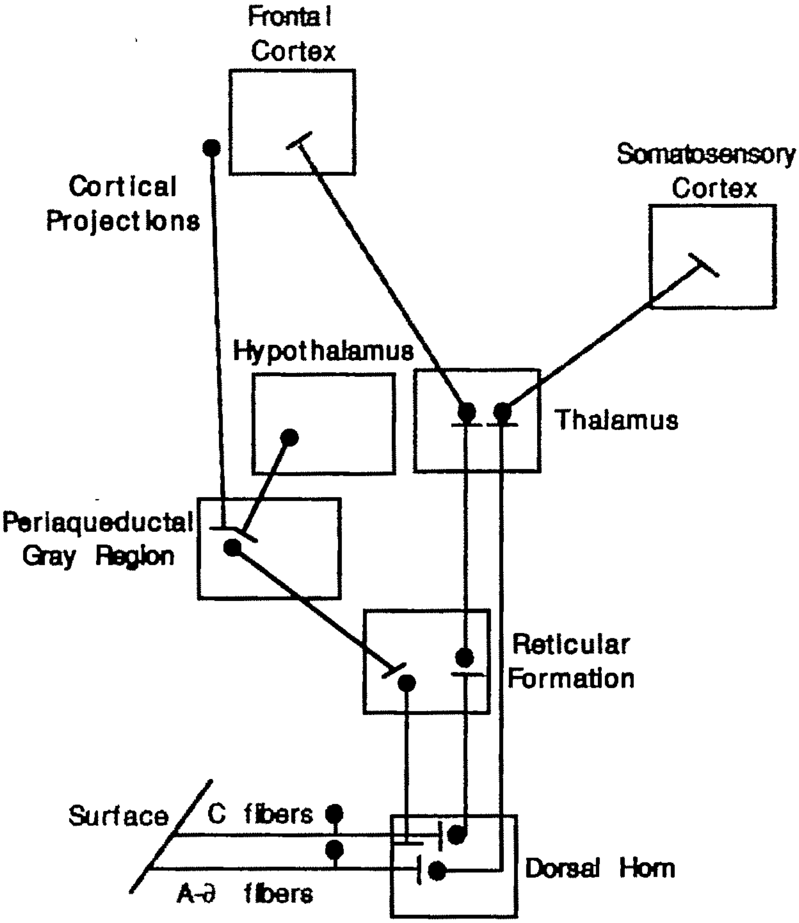


Figure 4. Diagram depicting the pain inhibitory system in contrast to the pain sensory system. Projections originating in the cortex and hypothalamus descend to the periaqueductal gray of the midbrain, the reticular formation, and then finally to the dorsal horn, where they inhibit the ascending nociceptive transmissions.

More important than being anatomically distinct and dissociable from nociception, the inhibitory streams are *teleologically* distinct from our PSS system as well. The inhibitory streams are not merely general purpose dampers; they are triggered by a very specific constellation of peripheral and cortical inputs. Otherwise, their removal would drastically increase our sensations of pain or make them all chronic. Such is not the case though. Lesioning the PAG does not appear to be tied to the onset of chronic pains.[73] Moreover, we do know that the inhibitory systems are activated (or inhibited) by nociception plus cortical arousal (stress) on the organism.[74] Fear and learned hopelessness also affect their activity.[75] Hence, the correlates for the firing of these neurons differ in kind from what our ascending pain system is sensitive to.

In addition, if we look at the connections of the inhibitory streams, we can see that they differ substantially from the PSS. It would not be proper to call them sensory systems or subsystems for they have no connections to the periphery. The pain inhibitory streams halt at the dorsal horn. Also unlike our ascending pain streams, the hypothamalus and dorsal raphe nuclei receive massive inputs from cortical processors, which presumably could carry information about our goals and immediate plans, what else is occurring in the environment, and our emotional context.[76] They have immediate access to information that the ascending pathways do not.

Finally, a two-system theory of pain sensation explains our evolutionary confusions. When we are under stress, it is often more adaptive not to feel pain than to be incapacitated by pain.[77] If we are fighting or fleeing from an enemy, it would be preferable to do so unencumbered by the need to nurse or protect our limbs, even if this results in more nursing or protecting later (when we are presumably safe). It is important to know when damage is occurring in our bodies, but it is equally important to be able to shut that information out when circumstances demand. A dual system would allow just such a contingency; we could inhibit our pains as needed, but then feel them again when the danger is gone. A PSS and a PIS then serve two different goals: the PSS keeps us informed regarding the status of our bodies. It monitors our tissues to maintain their intactness whenever possible. In contrast, the PIS shuts down the PSS when flight or fleeing is immanent, and then enhances the PSS response in moments of calm. If our brains are geared for motor control, then the dual pain system makes good biological sense.

---

[73] See Pagni.

[74] Fields and Basbaum; M.L. Mayer, "Periaqueductal Gray Neuronal Activity, Correlation with EEG Arousal Evoked by Noxious Stimuli in the Rat," *Neuroscience Letters*, xxviii (1982): 297-301.

[75] S.F. Maier, R.C. Drugan, J.W. Grau, "Controllability, Coping Behavior, and Stress-induced Analgesia in the Rat," *Pain*, xii (1982): 47-56; J.W. Lewis, J.T. Cannon, and J.C. Leibeskind, "Opioid and Non-opioid Synapse Mediates the Interaction of Spinal and Brain Stem Sites in Morphine Analgesia," *Brain Research*, ccxxxvi (1980): 85-91; J.W. Lewis, J.E. Sherman, and J.C. Leibeskind, "Opioid and Non-opioid Stress Analgesia, Assessment of Tolerance and Cross Tolerance with Morphine," *Journal of Neuroscience*, i (1981): 358-63; J.W. Lewis, M.G. Tordoff, J.E. Sherman, and J.C. Liebeskind, "Adrenal Medullary Enkephalin-like Peptides May Mediate Opioid Stress Analgesia," *Science*, ccxvii (1982): 557-59; Fields and Basbaum.

[76] See Kandel and Schwartz.

[77] R.J. Bodnar, "Effects of Opioid Peptides on Peripheral Stimulation and 'Stress'-Induced Analgesia in Animals," *Critical Reviews in Neurobiology*, vi (1990): 39-49; Kandel and Schwartz; J.L. Szekely, "Opioid Peptides and Stress," *Critical Reviews in Neurobiology*, vi (1990): 1-12.

In sum: the bottom-up and the top-down investigative strategies support the conclusion that our pain inhibitors form a separate and distinct system. They react to different types of stimuli. They have different points of input and output from the PSS, and their existence is supported by evolutionary considerations.

VII. REPRIEVE: THE SENSATION OF PAIN

If my story is correct, then, contrary to how things seem to us, the sensation of pain is not what is most important in pain processing. It is but one minor aspect of our entire pain and pain-inhibitory systems, which themselves are geared to help us flee, fight, or nurse ourselves, depending upon the circumstances. They function just as any of our perceptual systems do: they help us get around in our environment as effectively as possible. From an evolutionary perspective, visual sensations are not the *raison d'être* of our visual system, and auditory sensations are not the ultimate goal for our auditory system. Sensations of pain are no different. Insofar as the PSS is simpler than other perceptual systems, then it makes good sense to use pain as a paradigm case of our conscious experiences.

We must issue an important caveat, however: pains are the product of a complex sensory system that has to struggle to get itself heard. Our conscious sensations are the product of both nociception and activity of a PIS.[38] Even though nothing about painful experiences is deeply mysterious—they are not random, nor inexplicable, nor tied to our psychological whims—they still are poorly correlated with actual tissue damage. What is interesting and different about our pains is the PIS. It is geared to suppress or enhance the activity of a single sensory system, and there is no other system quite like it in our nervous system. We do have general inhibitory and excitatory systems in the brain, but none of these is specific to a particular sensory system. The PIS entails that our sensations of pain are almost independent of nociception. Hence, insofar as we have a PIS, then our sensations of pain are special and are not a typical example of a phenomenological process. Putnam and others are simply wrong to think of a pain as the exemplar of consciousness.

---

[38] This is not to say that the PIS is directly connected to our sensations of pain. It is not. We can remove bits and pieces of the PAG, for example, and still feel pain. Nevertheless, the PIS heavily influences when and where we feel pain.

VALERIE GRAY HARDCASTLE

Virginia Polytechnic Institute and State University

REFLECTIVE KNOWLEDGE IN THE BEST CIRCLES*

Is the existence of external things just an article of faith? Certainly not, says G. E. Moore,[1] who offers us a proof, thus aiming to remove Immanuel Kant's "scandal to philosophy."

> Moore's proof
> Here is a hand (a real, flesh and bone hand).
> Therefore, there is at least one external thing in existence.[2]

According to Moore, his argument meets three conditions for being a proof: first, the premise is different from the conclusion; second, he knows the premise to be the case; and, third, the conclusion follows deductively (*ibid.*, pp. 144-45). Further conditions may be required, but he evidently thinks his proof would satisfy these as well.

I. MOORE'S PROOF

As Moore is well aware, many philosophers will feel he has not given "any satisfactory proof of the point in question" (*ibid.*, p. 147). Some, he believes, will want the premise itself proved. But he does not try to prove it, and does not believe it can be proved. Proving that here is a hand requires proving one is awake, and this cannot be done.

Does Moore adequately answer the skeptic?[3] Many have denied it for the reason that he fails to rule out a crucial possibility: that our faculties are leading us astray—for example, that we are dreaming. Aware of this objection, Moore grants, in "Certainty,"[4] that to know, he is standing, he must know he is awake. The point "cuts both ways," however, and he would prefer to conclude that he does know he is awake since he does know he is standing. This has persuaded nearly nobody. On the contrary, some have thought him committed to an argument, *M* below, like the following:

> Argument *A*
> A1. This map is a good guide to this desert.
> A2. According to the map, an oasis lies ahead.
> A3. Therefore, an oasis lies ahead.

---

* I am pleased to acknowledge helpful discussion, first here at Brown University; at the University of California/Santa Barbara; at the Oxford Philosophical Society; at Cambridge University; and at the University of Granada.

[1] Proof of an External World," in his *Philosophical Papers* (New York: Collier, 1962), pp. 144-45. Originally in the *Proceedings of the British Academy*, xxv (1939): 273-300.

[2] This is, of course, a simplified version of Moore's proof.

[3] Henceforth 'knowledge' here will be short for 'plain knowledge', leaving aside Cartesian superknowledge.

[4] In his *Philosophical Papers*, pp. 226-51.

0022-362X/97/9408/410–30

Argument *M*

M1. My present experience is a veridical guide to reality (and I am not dreaming).

M2. My present experience is as if I have a hand before me.

M3. Therefore, here (before me) is a hand.

When challenged on premise A1, our desert dullard responds: "I must know A1, since the only way I could know A3 is through argument *A*, and I do know A3." Is this a just comparison? Is Moore's response to the skeptic relevantly similar?[5]

If Moore depends on argument *M* for his knowledge of M3, his response seems like the dullard's. The dullard is wrong to respond as he does. He must say how he knows his premise without presupposing that he already knows the conclusion. And Moore would seem comparably wrong in the analogous response to the skeptic. In explaining how he knows M1, he must not presuppose that he already knows M3.

Does Moore depend on argument *M* for his knowledge of M3? There is reason to think that he does not, given his emphatic acknowledgment that he cannot *prove* M3. After all, *M* would seem a proof of M3 just as good as Moore's own "proof of an external world." Moore concedes, in effect, that, if he does not know that he is not dreaming, *then* he does not know of the hand before him. But that is not necessarily because he takes himself to know M3 only through *M* or any other such argument. In any case, even if he is relying on some such argument, which would require making that concession, the defender of common sense has other options.

One might, after all, make that concession only because of the following "principle of exclusion" (PE):

> If one is to know that *h*, then one must exclude (rule out) every possibility that one knows to be incompatible with one's knowing that *h*.

As Moore grants explicitly, the possibility that he might be just dreaming is incompatible with his knowing (perceptually) that he has a hand before him. And this, in combination with PE, is quite sufficient to explain his concession above.

Suppose Moore is not depending on argument *M* for his knowledge of M3. Although he recognizes his need to know he is not dreaming, suppose that is only because he accepts PE, our principle of exclusion. Then the sort of ridicule cast on the dullard is misdirected against Moore. What is more, it is not even clear that Moore must know *how* he knows he is not dreaming if he is to know M3. That is not entailed by the application of the principle of exclusion. All that follows from such application is that Moore must know *that* he is not dreaming, not that he must know *how* he knows this.

---

[5] Compare Barry Stroud, *The Significance of Philosophical Scepticism* (New York: Oxford, 1984), chapters 1 and 3.

In fact, however, the historical Moore did rely on something very much like argument *M* (more on this below). So is he not, after all, exposed to the damaging comparison with the desert dullard? Not at all. There seems to be no good reason why, in responding to the skeptic, Moore must *show how* he knows he is not dreaming. Of course, his response to the skeptic would be enhanced if he could show that. But it now seems not properly subject to ridicule even if he is not then in a position to show how he knows he is not dreaming. The question he is addressing is *whether* he knows that he is not dreaming, and, at most, by extension, what grounds he might have for his answer to that question, in answering which he does not, nor need he, also answer the question of *how* he knows himself to be awake and not dreaming.

It might be replied that one cannot know that here is a hand if one's belief rests on the unproved assumption that one is awake. According to Moore, however, things that cannot be proved might still be known. Besides, even though he cannot prove that he is awake, he has "conclusive evidence" for it. Unfortunately, he cannot state his evidence, and the matter is left in this unsatisfactory state at the end of "Proof of an External World." But Moore has more to say in another paper of the period, "Four Forms of Scepticism."[6] There, he takes himself to know for sure about the hand before him, and takes this knowledge to be based on an inductive or analogical argument. We are told that introspective knowledge of one's own sensory experience, unlike perceptual knowledge of one's physical surroundings, can be immediate. While agreeing with Bertrand Russell that one cannot know *immediately* that one sees a hand, Moore thinks, contra Russell, that he can know it for certain. And he disagrees with Russell more specifically in allowing knowledge for certain about his hand through analogical or inductive reasoning from premises known introspectively.

It is doubtful, however, that any allowable form of inference—whether deductive, inductive, or analogical—will take us from the character of our experience to the sort of knowledge of our surroundings we ordinarily claim.

---

[6] In his *Philosophical Papers*, pp. 193-223.

Familiar skeptical scenarios—dreaming, evil demon, brain in a vat, and the like—show that our experience prompts but does not logically entail its corresponding perceptual beliefs. Experience as if there is a fire before us does not entail that there is a fire there, experience as if here is a hand does not entail that here is a hand, and so on. Perhaps what is required for one's beliefs and experiences to have certain contents entails that these could not possibly be *entirely* false or misleading. Indeed, some such conclusion follows from certain externalist and epistemic requirements on one's justified attribution of familiar contents to one's own experiences or beliefs. But even if that much is right—which is still controversial—one's experience or belief that here is a hand, or yonder a fire, might still be wildly off the mark. We cannot deduce much of our supposed knowledge of the external world from unaided premises about our experience.

As for inductive or analogical reasoning, only abductive reasoning—inference to the best explanation—offers much promise, but it seems questionable as a solution to our problem.[7] Suppose (a) that we restrict ourselves to data just about the qualitative character of our own sensory experience, and (b) that we view belief in a commonsensical external world as a theory postulated to explain the course of our experience. What exactly is the proposal? Is it proposed that when ordinarily we accept the presence of a hand before us, we do know, and know on the basis of an abductive inference? Or is it proposed, rather, that in such circumstances we have resources that would enable us to know if only we used those resources to make effective abductive arguments? The second, more modest, proposal is too modest, since it leaves our ordinary perceptual beliefs in a position like that of a theorem accepted through a guess or a blunder, one which we do have the resources to prove after much hard thought, but one which we have not come close to proving at the time when we are just guessing or blundering.

Even the modest proposal, moreover, seems unlikely to succeed. Could we form a rich enough set of beliefs purely about the qualitative character of our sensory experience, one rich enough to permit abductive inferences yielding our common-sense view of external reality?

---

[7] For Russell the "common-sense hypothesis" of independent physical objects is "simpler" than the supposition that life is but a dream (as he explains in *The Problems of Philosophy* (New York: Oxford, 1959), chapter 2). For W.V. Quine, the "hypothesis of ordinary physical objects" is "posited" or "projected" from the data provided by sensory stimulations: "Subtracting his cues from his world view, we get man's net contribution as the difference"—*Word and Object* (Cambridge: MIT, 1960), p. 5. That Quine's position is deeply problematic is shown by Stroud, chapter 6.

This seems doubtful when we consider (a) that such pure data beliefs could not already presuppose the external reality to be inferred, and (b) that the postulated common-sense "theory" of external reality must presumably meet constraints on abductive inference: for example, that the postulated theory be empirically testable and also simpler and less ad hoc than alternatives (for example, George Berkeley's). These requirements plausibly imply that our data must go beyond detached observations, and include some acceptable correlations. Yet these correlations are unavailable if we restrict ourselves to beliefs about the character of our experience.[8] Most especially are they unavailable, and most especially is the postulated inference implausible, when our database is restricted, as it is by Moore, to introspectively known facts of one's own then-present subjective experience and to directly recalled facts of one's own earlier experience. (If deprived of the epistemic resources of testimony and of retentive memory—except insofar as such resources can be validated by reason-cum-introspection, which is not very far if at all—then there is precious little we can any longer see ourselves as knowing, thus deprived.)

Accordingly, the skeptic has a powerful case against Moore's claim that our knowledge of the external world is based on an inductive or analogical inference from such information about our experience. It is not realistic to suppose that we consciously make such inferences in everyday life. It is less implausible to conceive of such inferences as implicit and/or dispositional, but even this strains belief. Besides, even granted that we make such inferences if only implicitly, do they yield simpler and less ad hoc hypotheses than alternatives? That is far from clear; nor do such hypotheses seem empirically testable and credible simply as explanations of the purely qualitative character of our then-present or directly recalled experience.

Having reached a dead end, let us have some second thoughts on Moore's view of perceptual beliefs as inferential, whereby he joined a venerable tradition, along with Russell himself. If perceptual knowledge is thus mediate and inferential, what knowledge can qualify as immediate and foundational? Modern philosophy begins with René Descartes's canonical answer to this question.[9]

---

[8] This is argued by Wilfrid Sellars in "Phenomenalism," in his *Science, Perception, and Reality* (New York: Routledge, 1963), pp. 60-105.

[9] The shift to discussion of Descartes may seem abrupt; however, what we find about the nature of immediate knowledge in that discussion has important implications for a position that Moore failed to explore. Skeptics who are willing to grant Descartes his immediate knowledge through introspection or rational intuition would need to explain exactly why perception could never yield such knowledge. (And what of memory?) The discussion of Descartes to follow is meant to highlight this issue.

II. DESCARTES'S CIRCLES

Descartes had two circles, not only the big famous one involving God as guarantor of our faculties, but also a smaller one found in the second paragraph of his third meditation, where he reasons like this:

> I am certain that I am a thinking being. Do I not therefore also know what is required for my being certain about anything? In this first item of knowledge there is simply a clear and distinct perception of what I am asserting; this would not be enough to make me certain of the truth of the matter if it could ever turn out that something which I perceived with such clarity and distinctness was false. So I now seem to be able to lay it down as a general rule that whatever I perceive very clearly and distinctly is true.[10]

Yet when he looks away from particular clear and distinct items, such as the proposition that he thinks, Descartes grants that a powerful enough being could deceive him even about what seems most manifest. Descartes grants that he could be astray in his beliefs as to what he perceives or remembers, and even in taking himself to intuit something as quite clear and distinct. This doubt must be blocked if one is to attain certainty by intuiting something as clear and distinct. Accordingly, Descartes launches the theological reflections that lead eventually to his nondeceiving God.

Even without the further boost of certainty provided by the proof of a nondeceiving God, however, Descartes takes himself to have attained some positive justification. Early in the third meditation, he takes himself to perceive clearly and distinctly that he thinks, which he takes to be what gives him the certainty that he thinks. He reasons that this clear and distinct perception would not give him such certainty if it were less than perfectly reliable, and apparently concludes from this that his clear and distinct perception is perfectly reliable. One could demand how he knows all these things: How can he be sure that he does clearly and distinctly perceive that he thinks, for one thing? How can he be sure that there is nothing else in his situation that could provide the degree of certainty involved? How can he be sure that the clarity and distinctness of his perception could not possibly provide that degree of certainty unless it were infallible? What could he say in response? Descartes might well have a uniform response to all such questions: in each case, he might appeal once again to clear and distinct perception, each of the things in question being something we are assured of by our clearly and distinctly perceiving it.

---

[10] *The Philosophical Writings of Descartes*, J. Cottingham, R. Stoothoff, and D. Murdoch, eds. (New York: Cambridge, 1975), Volume II, p. 24.

About the cogito, I wish to highlight not Descartes's answers to such questions, however, but the inference that he draws: "so I now seem to be able to lay it down as a general rule that whatever I perceive very clearly and distinctly is true." Just what is Descartes's argument in support of this general rule? Would his reasoning take the following form?

(1) Datum: I know with a high degree of certainty that I think.
(2) I clearly and distinctly perceive that I think, and that is the only, or anyhow the best account of the source of my knowledge that I think.
(3) So my clear and distinct perception that I think is what explains why or how it is that I know I think.
(4) But my clear and distinct perception could not serve as a source of that knowledge if it were not an infallibly reliable faculty.
(5) So, finally, my clear and distinct perception must be an infallibly reliable faculty.

The move from 1 and 2 to 3 is an inference to an explanatory account that one might accept for the coherence it gives to one's view of things in the domain involved. Elsewhere, Descartes does appeal to coherence at important junctures.[11] So he may be doing so here as well, although questions do arise about how Descartes views coherence. Does he accept the power of coherence to add justified certainty, and, in particular, would he claim infallibility for (sufficiently comprehensive and binding) coherence as he does for clear and distinct intuition?[12] In any case, the comprehensive coherence of his world view would be enhanced by an explanation of how clear and distinct perception comes to be so highly reliable, even infallible. And this is just what Descartes attempts, through his theological and other reasoning. Descartes can see that reason might take him to a position that is sufficiently comprehensive and interlocking—and thereby defensible against any foreseeable attack, no holds barred, against any specific doubt actually pressed or in the offing, no matter how slight. Unaided reason might take him to that position. Need he go any further? What is more: Might one reach a similar position while dispensing with the trappings of Cartesian rationalism?

---

[11] In his *Principles of Philosophy* (part IV, art. 205), for example, he notes that, if we can interpret a long stretch of otherwise undecipherable writing by supposing that it is written in "one-off natural language," where the alphabet has all been switched forward by one letter, and so on, then this is good reason for that interpretation. There, he also argues for his scientific account of reality in terms of certain principles by claiming that "it would hardly have been possible for so many items to fall into a coherent pattern if the original principles had been false"—*The Philosophical Writings of Descartes*, Volume I, p. 290.

[12] My attribution to Descartes is tentative because of the enormous bibliography on the "Cartesian Circle." In deference to that important tradition of scholarship, I do no more than suggest that there is logical space for an interpretation of Descartes that is perhaps more complex than many already tried, but that seems coherent and interesting. (I am myself convinced that this is Descartes's actual position, and defend this more fully in "How to Resolve the Pyrrhonian Problematic: A Lesson from Descartes," *Philosophical Studies*, LXXXV (1997): 229-49. In "Mythology of the Given," in *The History of Philosophy Quarterly* (forthcoming), I argue for the relevance of this Cartesian strategy to issues of empirical foundations that have divided philosophers since the Vienna Circle, and have pitted, for example, Wilfrid Sellars, Richard Rorty, and Donald Davidson, on one side, against C.I. Lewis, Carl Hempel, and Roderick Chisholm, on the other. My warm thanks to Lex Newman and James Van Cleve for helpful discussion of, and further references relevant to, this way of viewing Descartes.)

III. CIRCULAR EXTERNALISM

Compare now how Moore might have proceeded:

(1) Datum: I know with a high degree of certainty that here is a hand.
(2) I can see and feel that here is a hand, and that is the only, or anyhow the best account of the source of my knowledge that here is a hand.
(3) So my perception that here is a hand is what explains why or how it is that I know (with certainty) that here is a hand.
(4) But my perception could not serve as a source of that degree of justified certainty if it were not a reliable faculty.[13]
(5) So, finally, my perception must be a reliable faculty.

Moore could, of course, go on to say more about the nature of the perception that assures him about the hand. He might still say that such perception involves an implicit inference from what is known immediately and introspectively, perhaps an inductive or analogical inference of some sort. That might make his view more comprehensively coherent, but we have already seen reasons why postulating such an inference is questionable. So we focus rather on a second alternative: Moore might well take perceiving to involve no inference at all, not even implicit inference, but only transfer of light, nerve impulses, and so on, in such a way that the character of one's surroundings has a distinctive impact on oneself and occasions corresponding and reliable beliefs. This might also amount eventually to a comprehensively coherent view of one's knowledge of the external world. Its epistemologically significant features would not distinguish it in any fundamental respect from the procedure followed by Descartes.

The theme of accidentally true belief has loomed large in the epistemology of recent decades. The Gettier problem, for example, is posed by a justified belief true for reasons far removed from whatever causes it to be held and justified. Externalist conceptions of propositional knowledge focus on this theme, as do one offered by Peter Unger (nonaccidentally true belief) and one offered by Alvin Goldman (belief caused by the truth of its content). Robert Nozick's tracking account is also a conception of this sort: S knows that p if and only if S believes correctly that p, and also (in the circumstances): *both* it would have been true that p only if S had believed it, *and* if it had not been true that p, then S would not have believed it.[14]

---

[13] Here, one would reduce Descartes's requirement of infallible certainty.

Why are these conceptions of knowledge of special interest to us here? Because each offers a way to explain how one can know that p without reasoning from prior knowledge. The key idea exploited is this: you can know something noninferentially so long as it is no accident or coincidence that you are right.

Both the tracking and the causal accounts defensibly require a special nonaccidental connection between the belief and the fact believed. Nevertheless, in each case other levels of accidentality remain. Suppose I fancy myself a connoisseur of tomato ripeness, but suffer from a rare form of color blindness that precludes my discerning nearly any shade of red except that displayed by this particular tomato. Therefore, my judgments of tomato ripeness are in general apt to be right with no better than even chance. But when it is the particular (and rare) shade of red now displayed, then I am nearly infallible. Oblivious to my affliction, however, I issue judgments of tomato ripeness with abandon over a wide spectrum of shades of red. Assuming that, unknown to me, the variety of tomato involved always ripens with this shade of red, then my belief that this tomato is ripe is in step with the truth, and arguably satisfies the requirements of Unger, Goldman, and Nozick. But, again, it is nevertheless in some relevant sense or respect only an accident that I am right in my belief.[15] We need a clearer and more comprehensive view of the respects in which one's belief must be nonaccidentally true if it is to constitute knowledge.

Unaided, the tracking or causal requirements proposed suffer from a sort of tunnel vision. They permit too narrow a focus on the particular target belief and its causal or counterfactual relation to the truth of its content. Just widening our focus will not do, however, if we widen it only far enough to include the process that yields the belief involved. We need an even broader view.

---

[14] Unger, "An Analysis of Factual Knowledge," this JOURNAL, LXV, 6 (March 21, 1968): 157-70; Goldman, "A Causal Theory of Knowing," this JOURNAL, LXIV, 12 (June 22, 1967): 357-72; Nozick, *Philosophical Explanations* (Cambridge: Harvard, 1981), chapter 3.

[15] For an early statement of this sort of problem, urged against Nozickian tracking, see Colin McGinn's "The Concept of Knowledge," *Midwest Studies in Philosophy*, IX (1984): 529-54.

IV. VIRTUE EPISTEMOLOGY

> [When]...thought is concerned with study, not with action or production, its good or bad state consists [simply] in being true or false. For truth is the function of whatever thinks....
>
> Hence the function of each of the understanding parts is truth; and so the virtue of each part will be the state that makes that part grasp the truth most of all. —Aristotle[16]

Virtue epistemology is distinguished by its emphasis on the subject as seat of justification. In order to qualify as knowledge, a belief must be "apt," epistemically so, in a strong sense that goes beyond its being just a belief that coheres well within the subject's perspective. The "tracking account" (Nozick) sees here little more than a claim about that belief's counterfactual relation to the truth of what is believed. "Reliable-indicator" accounts require rather that the belief itself or the reasons for it have properties nomically sufficient for its truth (David Armstrong, Marshall Swain[17]). "Reliable-process" accounts focus instead on the cognitive process, beneath the skin, that yields the belief, and on the truth ratio in the products of that process, actual and counterfactual (Goldman).

It is rather the subject and her cognitive virtues or aptitudes which hold primary interest for virtue epistemology. Consider the athletic virtues of a tennis champion. When we say that a shot is not just a winning shot but a skillful one, we imply a comment on the player as shotmaker. Suppose a tyro wields a racquet on a court and, unaware of the approaching ball, issues what amounts by luck to a stylish and effective backhand stroke. Such a shot might be an unreturnable winning shot, but it would manifest no real skill.

Why are we unwilling to admire a performance as "skillful" if it manifests only a fleeting, or even an instantaneous state of the agent's? Skills, abilities, competences, aptitudes, prowess—these come and go, true enough, but they do not flit by instantaneously. Why not? Why do we tend to define these concepts so as to require such stability? We might have defined similar concepts without requiring stability. Why do we define these concepts as we do? Why have we adopted these and not others? Should one not expect that, other things being equal, the more clearly useful a concept is to us, the more likely it is that we shall retain it? People need to know who are dependable members of their group—this is a *kind* of thing we need to monitor in a great variety of contexts, with a great variety of objectives. Cooperative success depends on the group's ability to monitor people's aptitudes and ineptitudes. So it is no surprise that the sorts of aptitudes (skills, competences, virtues) that we recognize and admire are those which linger stably.[18]

---

[16] *Nichomachean Ethics*, T. Irwin, trans. (Indianapolis: Hackett, 1985), 1139a27-30.

[17] Armstrong, *Belief, Truth, and Knowledge* (New York: Cambridge, 1973); Swain, *Reasons and Knowledge* (Ithaca: Cornell, 1981).

To praise a performance as skillful or an action as right, or a judgment as wise or apt, accordingly, is to assess not only the action or the judgment, but also the reflected aptitude or character or intelligence. This is a distinctive view with versions both in epistemology and in ethics. It is distinctive in that the rightness of an action (or a choice) and the aptness (or positive epistemic status) of a belief would involve not just whether the performance is *optimific* (if an action) or *true* (if a belief); nor just whether a good enough procedure was followed, perhaps accidentally, in arriving at that choice, or whether a good enough cognitive process chanced to lead to the belief; nor even just whether a rule, in effect, somehow demands that choice or that belief in those circumstances. Our virtue epistemology and virtue ethics focus rather on the agent and cognizer. When the agent's actions are said to be *right* and the cognizer's beliefs *knowledge*, we speak implicitly of the virtues, practical or intellectual, seated in that subject, which (a) give rise to that action or belief, adding to the subject's worth as agent or cognizer, and (b) make him reliable and trustworthy over an interesting spread of possible choices or beliefs, and circumstances.

V. VIRTUE AND COHERENCE

Can we explain what distinguishes a system of beliefs (and experiences) that, internally regarded, is intellectually virtuous and admirable? Presumably, our explanation would involve the system's explanatory coherence, overall simplicity, lack of adhoc epicycles, and so on. Philosophical mythology contains creatures who excel in all such respects, however, though their beliefs fall short of being knowledge even when true. Take the brain in a vat, for example, or the victim of the evil demon. An adult recently envatted, or victimized by the demon, can be indistinguishable from the best of us in respect of the comprehensiveness and coherence of their beliefs and experiences. Even when right about environing objects and events, such a victim's beliefs are far from being knowledge.

---

[18] The social utility of concepts is invoked occasionally to defend a proposed account of knowledge in my *Knowledge in Perspective* (New York: Cambridge, 1991); see pp. 27, 275. The form of argument involved is used insightfully in E.J. Craig's *Knowledge and the State of Nature* (New York: Oxford, 1990), where it takes center stage.

Might such comprehensive internal coherence exhaust all cognitive or intellectual virtue, at least when it comprises not only beliefs but also experiences? To assent here would be overbold even for a rationalist. Yet coherence is, of course, valued not only by philosophers but by the reflective more generally. One also wants faculties and virtues beyond reflective, coherence-seeking reason: perception, for example, and memory. Equally, internal coherence goes beyond such faculties, and requires reason, which counts for a lot in its own right. But why should that be so, if comprehensive coherence is no guarantee of truth, if the internal coherence enjoyed by the envatted yields little if any truth?

Compare first reason with memory. Input beliefs are required for retentive memory and inferential reason, which then yield beliefs as outputs. Retentive memory yields again the input belief itself, while inferential reason yields a new belief. Even the most excellent transmission faculty will not guarantee the truth of its output, which will depend not only on the transmission but also on the inputs. But our transmission faculties are valuable even so, if only because they combine with other faculties to increase vastly the total yield of true beliefs.

How does internal coherence, of little significant epistemic value in itself, become more valuable when combined with external aptness? Coherence-seeking inferential reason, like retentive memory, is of epistemic value when combined with externally apt faculties of perception, because when so combined it, like retentive memory, gives us a more comprehensive grasp of the truth than we would have in its absence.

Good perception is in part constituted by certain transitions from experiences to corresponding beliefs—as is the transition from the visual experience characteristic of a tomato seen in good light to belief in the tomato. Other such transitions help constitute good introspection, as when one's headache prompts awareness of it as a headache. Finally, if the comprehensive coherence of one's system of beliefs is at least in part responsible for its constitution and persistence, it thereby manifests a virtuous faculty of reason. Such comprehensive coherence is not just mechanical, but must reflect appropriate sensitivity to factors like adhocness, simplicity, and explanatory power. And it must include, not only belief/belief connections, but also experience/belief connections constitutive of good perception, and conscious-state/belief connections constitutive of good introspection. This broader conception of the coherence of one's mind involves not only the logical, probabilistic, and explanatory relations among one's first-order beliefs, but also coherence between these beliefs and one's sensory and other experiences, as well as comprehensive coherence between first-order experiences, beliefs, and other mental states, on one side, and beliefs about first-order states, on the other.

We may well ask about certain aspects of broad coherence—for example, the experience/belief transitions as well as the enumerative and abductive inferences involved—why these should be viewed as adding to the subject's intellectual worth or merit. "Because it is truth conducive," or at least in good measure for that reason, we are told, "because it increases the likelihood that the subject will have true beliefs and avoid false ones." But that is obviously false of victims in skeptical scenarios, who nevertheless are internally coherent, and even epistemically justified.

Although that seems undeniable, we can perhaps understand it comfortably if we distinguish two sorts of epistemic justification: (a) *S* is "same-world justified" in believing *P* in world *W* if and only if *S* believes *P* in *W* in virtue of a faculty that in *W* is truth conducive; and (b) *S* is "actual-world justified" in believing *P* in world *W* if and only if *S* believes *P* in *W* in virtue of a faculty that in our actual world is truth conducive.

Such relativizing and contextualizing is familiar enough in ordinary thought and speech. Here, it enables us to combine the following theses: (a) our broad coherence is necessary for the kind of reflective knowledge traditionally desired; and (b) such broadly coherent knowledge is desirable because in our actual world it helps us approach the truth and avoid error. This is not to deny that there is a kind of "animal knowledge" untouched by broad coherence. It is rather only to affirm that beyond "animal knowledge" there is a better knowledge. This reflective knowledge does require broad coherence, including one's ability to place one's first-level knowledge in epistemic perspective. But why aspire to any such thing? What is so desirable, epistemically, about broad coherence? Broad coherence is desirable because it yields integrated understanding, and also because it is truth conducive, even if in a demon world broad coherence fails this test and is not truth conducive. Even so, we can still regard broad coherence as intellectually valuable and admirable so long as we do not regard our world as such a world.

We are now in just the position of arch-internalist Descartes. Consider the following passage:

> The fact that an atheist can be "clearly aware that the three angles of a triangle are equal to two right angles" is something I do not dispute. But I maintain that this awareness of his [*cognitionem*] is not true knowledge [*scientia*], since no act of awareness that can be rendered doubtful seems fit to be called knowledge [*scientia*]. Now since we are supposing that this individual is an atheist, he cannot be certain that he is not being deceived on matters which seem to him to be very evident (as I fully explained). And although this doubt may not occur to him, it can still crop up if someone else raises the point or if he looks into the matter himself. So he will never be free of this doubt until he acknowledges that God exists.[19]

Descartes considers reasons to doubt, not only one's faculties of perception, memory, and introspection, but even one's faculty of intuitive reason, by which one might know that 3+2=5, that if one thinks one exists, and the like. He defends against such doubts by coherence-inducing theological reasoning that yields an epistemic perspective on himself and his world, in terms of which he can feel confident about the reliability of his faculties, including the very faculties employed in arriving, via a priori theological reasoning, at that perspective on himself and his world, the perspective that enables him to see his world as epistemically propitious.[20]

In structure, virtue perspectivism is thus Cartesian, though in content it is not. Radical rationalism admits only (rational) intuition and deduction (along with memory) as its faculties of choice (or anyhow of top choice) and wishes to validate all knowledge in terms of these faculties; thus the Cartesian grand project. Virtue perspectivism admits also perception and introspection, along with intuition and deduction, as well as inductive and abductive reasoning. Gladly using all such faculties, it also accepts through testimony the aid of one's epistemic community. Fortunately, the overview thus attained inspires confidence in the means used.

Rejected as viciously circular by Descartes's critics, and by many today, our procedure does present a troubling aspect of circularity. A closer look, however, may show this to be only an illusion.[21]

---

[19] This passage is from the Second Set of Replies as it appears in *The Philosophical Writings of Descartes*, Volume II, p. 101. I must add, however, that where this translation says that an atheist can be "clearly aware," Descartes's Latin is *clare cognoscere*.

[20] Although unremarked by Descartes, the role of dreams in his perception skepticism is analogous to a role assignable to paradoxes and aporias in a parallel skepticism vis-à-vis rational intuition.

[21] See my "Philosophical Skepticism and Epistemic Circularity," *Proceedings of the Aristotelian Society*, Supplementary Volume LXIV (1994): 263-90; and compare Stroud's response, "Scepticism, 'Externalism', and the Goal of Epistemology," pp. 291-307.

VI. EPISTEMIC CIRCULARITY: WHAT IS THE PROBLEM?

"I think, therefore I am," says Descartes, adding: "Here at last is something I really know. But what is it about this knowledge that makes it knowledge? As far as I can see, it is knowledge because it is a clear and distinct intuition. But it would not be real knowledge unless such intuition were reliable. So I can already lay it down as a general rule that clear and distinct intuition *is* reliable."

"Here is a hand," says Moore, adding: "Here is something I really know. But what gives me this knowledge? As far as I can see, it is knowledge in virtue of being a deliverance of perceptual experience. But it would not be knowledge if I were dreaming. So I can already conclude that I am *not* dreaming."

Descartes goes on to buttress the reliability of his rational intuition by developing a theology through vigorous use of that very rational intuition. Moore can similarly appeal to what he knows about his reliable senses on the basis largely of those very senses.

But is not any such reasoning circular? Yes, circular it does seem to be, "*epistemically* circular," let us say. But is it *viciously* circular? Skeptics through the ages have attacked it as such. Sextus Empiricus already uses the tropes of Agrippa in order to develop the so-called *diallelus*, or "problem of the criterion." Many have followed his lead in a long tradition. Today, skepticism cum relativism has spread beyond epistemology and ethics, beyond philosophy, and even beyond the academy, and its champions often wield circularity as a weapon. But, again: Is such circularity vicious? To say that it is vicious, in the present context, is to say that it is somehow bad, intellectually bad, that it puts us in a situation that is somehow intellectually unsatisfactory. When we ask how the circularity is vicious, therefore, what we want to know is just how it puts us in an unsatisfactory state: When we reason in the way alleged to be viciously circular, wherein lies the defect in our reasoning or in the resulting state?

Largely through the use of rational intuition, Descartes supports the view that rational intuition is reliable and that through its exercise he knows that he thinks and exists. Largely through the use of perception, Moore could support the view that perception is reliable, that he is not misled by a dream, and that through the exercise of perception he knows of the hand before him. If a crystal ball claims itself to be reliable, then, largely through the crystal ball, a crystal-ball gazer could support the view that such gazing is reliable, that it is rarely misleading, and that through the crystal ball he can foretell the future.

Epistemic circularity is vicious, it might be said, because it would make the gazer as well justified as Descartes or Moore. *Since* there is no way to support adequately the view that intuition is reliable, or that perception is reliable, without employing those very faculties; and *since* the same goes for memory, deduction, abduction, and testimony; *therefore*, there is no way to arrive at an acceptable theory of our knowledge and its general sources.

Perhaps that shows only how defective is the attempt to develop such a general theory of one's knowledge and its sources. There is an easy way to avoid the intellectual discomfort of having to use a faculty in answering the question whether that faculty is reliable, namely, not to ask the question. Call this the *avoidance strategy*.

Of course, we shall hardly lack company if we avoid philosophy because we find it frustrating. But the avoidance strategy that I wish to consider is not just a rejection of what seems too difficult for one's own intelligence. The implication of the avoidance strategy is not that there is something lacking in one's intelligence but that there is something wrong with the questions avoided.

Much might indeed be wrong with our very general, philosophical questions. Many find them too abstract, too impractical, too useless, and so on. But these are not the concerns of my avoidance strategist. He is, after all, a philosopher. His concern is not that the questions are just too hard for his intelligence, nor is it their abstractness, impracticality, or uselessness. He would hardly have gone into philosophy with such concerns, nor are they his concerns now. Difficulty, abstractness, impracticality, and uselessness are not in his view disqualifying drawbacks.

Why then should one as philosopher avoid questions of epistemology, such as those about the reliability of one's faculties? These questions become pressing with the realization that only if they reliably yield truth can our faculties yield knowledge. This is not just a commitment peculiar to contemporary reliabilism. Indeed, it is found already in Descartes, who, as we have seen, also stresses that intuition (and clear and distinct perception) yields knowledge only if reliable.

Consider again our principle of exclusion:

> If one is to know that *p*, then one must exclude (rule out) every possibility that one knows to be incompatible with one's knowing that *p*.[22]
> (By 'excluding' here I mean 'knowing not to be the case'.)

---

[22] Chapter 1 of Stroud's *The Significance of Philosophical Scepticism* is an illuminating discussion of this principle and its importance for understanding philosophical skepticism. In "How to Resolve the Pyrrhonian Problematic: A Lesson from Descartes," I suggest how to derive it from other principles with independent plausibility.

On the basis of PE, we can see that in order to know that *p*, one must know that the faculties employed in arriving at one's belief that *p* are reliable faculties. After all, just consider the possibility that one's operative faculties were unreliable. That is surely a possibility generally known to be incompatible with attaining knowledge through them. Unreliable mechanisms of belief acquisition will not yield knowledge.

If the principle of exclusion is right, therefore, one cannot possibly know that *p* unless one knows that the faculties involved are reliable. But this is just the sort of knowledge that we seem able to attain only through epistemically circular reasoning.

One might, of course, question the principle of exclusion.[23] One might hold that in order to know that *p*, one's pertinent faculties need only *be* reliable; one need not *know* them to be reliable.

One might, for example, appeal to a conception of knowledge as mere tracking. One might grant Richard Rorty that causation should not be confused with justification, while joining Nozick in taking tracking as the essence of knowledge. "To know is just to mirror (or to track) nature. Justification is quite another matter. Justification of some sort may well require the principle of exclusion. Thus, it may be that in order to be justified in believing that *p*, one must exclude every possibility one knows to be incompatible with one's knowing that *p*. But such justification is *not* required for simple knowledge."

That response seems essentially right. What is more, even Descartes would agree. For Descartes, you will recall, our knowledge that our faculties are reliable, even our faculty of reason, depends on our knowledge of God's epistemic good will. Yet, as we have seen, Descartes grants explicitly that the atheist mathematician can know some mathematics.

The knowledge of an atheist is said to be *cognitio*, however, a second-class accomplishment by comparison with *scientia*. *Scientia*, by contrast, does require relevant knowledge of one's reliability. Only thus can one repel doubts about the possible unreliability of one's faculties. Only thus can one exclude a possibility evidently incompatible with one's knowing that *p*, namely, the possibility that only unreliable faculties yield one's belief.

---

[23] For one thing, as it stands it leads, apparently, to a vicious regress. But that is an illusion. After all, what PE requires one to rule out is, not every possibility incompatible with one's knowing that *p*, but rather every possibility known to be thus incompatible. Since, for one thing, knowledge requires belief, the regress is hence not infinite, nor does it seem vicious.

By analogy, we can more generally distinguish *animal* knowledge, which requires only that one track reality, on one hand, and *reflective* knowledge, on the other, which in addition requires awareness of how one knows, in a way that precludes the unreliability of one's faculties. Unlike Descartes's *cognitio* and *scientia*, our more general animal and reflective knowledge do not require infallible reliability, but only a high level of reliability.[14] The avoidance strategy now has not only the cost of suppressing philosophical curiosity about knowledge. We can now see how it also precludes first-level reflective knowledge, and of course *scientia*.

Given these costs, what again counts in favor of avoidance? So far we have been told that we must avoid epistemic circularity because it entails arriving at a generally positive view of one's faculties only by use of those very faculties. But why should that be frustrating when it is the inevitable consequence of its generality? So far the answer is only that the superstitious crystal gazer could reason analogously and with equal justification in defense of his own perspective. How damaging is this?

Suppose we grant the gazer epistemic justification and internal coherence equal to our own. Still, internal coherence is clearly insufficient. Is that not obvious in view of paranoia, hypochondria, and similar psychoses? Logical brilliance permits logical coherence but does not even ensure sanity, much less general epistemic aptitude. There are faculties other than reason whose apt functioning is also crucial to the subject's epistemic welfare.

In light of that result, why not distinguish between the gazers and the perceivers in that, though both reason properly and attain thereby coherence and justification, only the perceivers' beliefs are epistemically apt and constitute knowledge?

On this view, the crystal gazers differ from the perceivers in that gazing is not reliable while perceiving is. So the theory of knowledge of the perceivers is right, that of the gazers wrong. Moreover, the perceivers can know their theory to be right when they know it in large part through perception, since their theory is right and perception can thus serve as a source of knowledge. The gazers are, by hypothesis, in a very different position. Gazing, being unreliable, cannot serve as a source of knowledge. So the perceivers have a good source or basis for their knowledge, but the gazers, lacking any such source or basis, lack knowledge.

---

[14] This distinction figures in my *Knowledge in Perspective* (New York: Cambridge, 1991), pp. 240, 282.

Still one might insist that the perceivers should not be so smug. They should still feel acute discomfort and intellectual frustration. This I find a very widely shared view, in epistemology and, mutatis mutandis, far beyond. According to Barry Stroud,[25] the perceivers can at best reach a position where they can affirm the conditional proposition that if their perception is reliable, then they know. And he has recently reemphasized what is essentially the same thesis as follows:

> ...Sosa's "externalist" could say at most: "If the theory I hold is true, I do know or have good reason to believe that I know or have good reason to believe it, and I do understand how I know the things I do." I think...we can see a way in which the satisfaction the theorist seeks in understanding his knowledge still eludes him. Given that all of his knowledge of the world is in question, he will still find himself able to say only "I might understand my knowledge, I might not. Whether I do or not all depends on how things in fact are in the world I think I've got knowledge of."[26]

It is not easy to understand this position, however. If our perceivers believe (a) that their perception, if reliable, yields them knowledge, and (b) that their perception is reliable, then why are they restricted to affirming only the conditional, *a*, and not its antecedent, *b*? Why must they wonder whether they understand their relevant knowledge? Indeed, to the extent that they are really convinced of both *a* and *b*, it would seem that, far from being logically constrained *to* wondering whether they know, they are, on the contrary, logically constrained *from* so wondering. After all, first, if you are really certain that *p*, then you cannot well consider whether you know it without thinking that you do. Moreover, second, is it not incoherent to be convinced that *p* and yet wonder whether *p*?

In sum, I see no sufficient reason to settle either for irresoluble frustration or for the avoidance strategy. The main argument we have seen for that depends on the claim that, if we allow the circular defense offered for externalist epistemology, then the gazers turn out no less epistemically justified than the perceivers. In a sense that is true: but then in a sense they are equally internally justified, equally coherent. Nevertheless, their beliefs are *not* equally apt in all epistemically relevant respects. Perception is, of course, reliable while gazing is not. Therefore, the perceivers are right and apt both in their particular perceptual beliefs, at least generally, and in their theory of knowledge—for it all rests in large measure on their reliable perception. By contrast, the gazers are wrong and inapt both in their particular gaze-derived beliefs and in their theory of knowledge—for it all rests on their unreliable gazing. Moreover, I see no reason why the perceivers must be restricted to affirming only the conditional that, if perception is reliable, then they know. I see no reason why they cannot also affirm the antecedent, why they cannot believe, both rationally and aptly, that perception is reliable and does enable them to know.

---

[25] "Understanding Human Knowledge in General," in M. Clay and K. Lehrer, eds., *Knowledge and Scepticism* (Boulder: Westview, 1989), pp. 31-50, here p. 47.

[26] "Scepticism, 'Externalism', and the Goal of Epistemology," pp. 303-04.

VII. CIRCLES BEYOND BELIEF

Why require the appeal to comprehensive enough coherence for justification, an appeal that I have attributed, tentatively, to Descartes, as part of what justifies his recourse to theology in accounting for true knowledge (*scientia*)? Why not say that what justifies is that one's beliefs be caused by the gods? If the question arises—why not add that *this* belief itself is justified because it is itself caused by the gods?—we could, of course, proceed in this simplified way without worrying about coherence or about the source of these beliefs beyond attributing them to divine agency. But that is not the way we are built, most of us: we just do not acquire such beliefs the way we do acquire beliefs willy-nilly when we open our eyes in good light. But what if we were built that way? Would we then be justified in having such beliefs, and in explaining our justification for having them, by their origin in divine agency? Would we then be justified to the degree and in the way in which Descartes is justified or in the way in which our imagined Moore would be justified through his appeal to a more ordinary reliabilism than that of Descartes? Internally regarded, the structure of beliefs would share prominent features in all three cases. Of course, from our Moorean, common-sense position we can object both to Cartesianism and to the invocation of the gods. These views are internally coherent, but we might still reject them as wrong. And we might be able to explain what is wrong with them, from our point of view, especially if our point of view rules out their leading ideas. But they can, for their part, return the favor. Besides, we can anyhow imagine someone brilliant but insane, who weaves a system of immense interlocking complexity, but one wholly detached from reality as we know it commonsensically. Such a madman could object to our common-sense beliefs in a way that would seem relevantly analogous to the way in which we would object to his mad beliefs.

What all of that shows, it seems to me, is nothing more than that knowledge does not live by coherence and truth alone. Knowledge requires truth and coherence, true enough, but it often requires more: for example, that one be adequately related, causally or counterfactually, to the objects of one's knowledge, which is not necessarily ensured by the mere truth-*cum*-coherence of one's beliefs, no matter how comprehensive the coherence. Madmen can be richly, brilliantly coherent; not just imaginary madmen, but real ones, some of them locked up in asylums. Knowledge requires not only internal justification or coherence or rationality, but also external warrant or aptness. We must be both in good internal order and in appropriate relation to the external world.

ERNEST SOSA

Brown University

BOOK REVIEWS

Poetic Justice: The Literary Imagination and Public Life. MARTHA C. NUSSBAUM. Boston: Beacon Press, 1995. xix + 143p. Cloth $20.00.

Poetic Justice is a self-consciously public book. Its main thesis is: "The ability to think of people's lives in the novelist's way is...an important part of the equipment of a judge" (99) and of citizens generally. Here, Martha Nussbaum echoes John Dewey[1]: "Imagination is the chief instrument of the good. It is more or less a commonplace to say that a person's ideas and treatment of his fellows are dependent upon his power to put himself imaginatively in their place" (*ibid*, p. 348).

The chief public problems whose sound treatment calls for novelistic thinking are, according to Nussbaum, racism and socioeconomic inequality. In order to address these problems effectively, judges and ordinary citizens alike "must...be capable of fancy and sympathy. They must educate not only their technical capacities but also their capacity for humanity" (121). Again, the similarity to Dewey is marked: "Wherever social divisions and barriers exist, practices and ideas that correspond to them fix metes and bounds, so that liberal action is placed under restraint.... Art is a mode of prediction not found in charts and statistics, and it insinuates possibilities of human relations not to be found in rule and precept, admonition and administration" (*ibid.*, pp. 348-49).

Whether the case for the role of imagination in encouraging liberal and humane public thinking is a strong one will depend for many readers on their attitudes toward liberal thinking. *Poetic Justice* has already been received in a number of public magazines as a strongly partisan work, as indeed it asks to be. But the book not only exhorts, it also argues, making its case in four interrelated chapters which treat, successively, the nature of literary imagination, problems with utilitarian reasoning, the rationality of emotions, and the role of novelistic imagination in jurisprudence.

The chapter on literary imagination argues that the novel "leads [readers] into certain postures of the mind and heart and not others" (2). Novelistic thinking illuminates both the nature of human well-being and the nature of the rational, deliberative processes of a good judge. This is so because the novel "takes as it theme, we might

---

[1] *Art as Experience* (New York: Minton, Balch, 1934).

0022-362X/97/9408/431–34

say, the interaction between general human aspirations and particular forms of social life that either enable or impede those aspirations, shaping them powerfully in the process" (7). Along with both economic analysis and moral and political theory, novels function as central vehicles for our asking, in a variant of former New York Mayor Ed Koch's famous query: "How are we doing, various of us in various ways?" Again, echoes of Dewey: "Literature conveys the meaning of the past that is significant in present experience and is prophetic of the larger movement of the future. Only imaginative vision elicits the possibilities that are interwoven within the texture of the actual" (*ibid.*, p. 345).

These general claims are then cogently illustrated in a chapter on Charles Dickens's *Hard Times*, in which Thomas Gradgrind is taken, by Nussbaum following Dickens, as illustrating the failures of utilitarian thinking alone in forming both personal and public policy. The list of failures is a standard roster of charges. Utilitarianism misconstrues all goods as commensurable; it focuses on the aggregate amount of good in the world at the expense of due recognition of the separateness and importance of distinct human lives; it misconstrues all human behavior as aimed at maximizing satisfactions; it misconstrues preferences themselves as simply given: exogenous to social choices and largely immune to recasting through moral or imaginative thinking that is not calculative. More interesting than this roster of charges, however, is the account of what Dickens, through his language and through some of his characters, sees about human life. Individuals have depth and particularity. Any human life is formed through the bearing of common problems and hopes in particular circumstances. Agency and freedom to choose the shape of one's life matter to us, over and above any quantum of satisfactions (27-29). Above all, *Hard Times* cultivates in us, its readers, "the ability to see one thing as another, to see one thing in another" (36)—for example, to see *that* as sympathy or *that* as callousness. Such "fancy" or "metaphorical imagination" is "a morally crucial ability, without which both personal and social relations are impoverished" (37). Without it, we become unable significantly to ask whether laws or actions or characters are just or charitable or mean spirited. Without it, we cannot learn to "cherish things for their own sake" (42) or to appreciate "the equality and dignity of all human life" (46).

Apt employments of such concepts not only rest on metaphorical imagination; they are also emotion saturated. It is difficult here to part understanding from feeling. Hence it is a mistake to construe emotions—for example, indignation at injustice—as blind forces, as involving judgments that are necessarily false, as irrelevant to public life, or as inherently aimed only at particular individuals. But whose emotions are genuinely perceptive and trustworthy? Here, Nussbaum appeals to Adam Smith's account of the judicious spectator as a concerned but not personally involved friend—disinterested, empathetic, imaginative, and in possession of wide factual knowledge. Reading novels, she suggests, helps to make us such spectators. "Literary readership...offers an artifical construction of the position of such a spectator" (72). By entering into this artificially constructed position through reading, we become better judicious spectators and judges in our personal and political lives.

In her final chapter, specifically on jurisprudence, Nussbaum shows, through the criticism of pieces of jurisprudential reasoning in actual cases, how to put metaphorical imagination judiciously to work, along with factual knowledge and concern for principles, in a public, institutional setting. The judge, along with modern readers in general, is involved in a "search for a human good that we are trying to bring about in and for the human community"—democratic equality (83-84). Some judges and legal reasoners search better or worse than others, depending on whether they bring metaphorical imagination and narratively motivated sympathies into play in their reasoning or not. Herbert Wechsler, in a famous article criticizing the holding in *Brown v. Board of Education* in the name of abstract fairness, errs by maintaining "a lofty distance from the human facts of the matter" of segregation (89). In contrast, Justice John Paul Stevens in his dissenting opinion in *Hudson v. Palmer* "does imagine...[the] legitimate interest in his property" (103) of the prisoner Palmer; and Richard Posner, in *Carr v. Allison Gas Turbine*, manifests literary skill in appreciating the "human meaning" (105) of a series of incidents that he then holds to constitute sexual harassment. Finally, Nussbaum criticizes the majority opinion upholding Georgia's anti-sodomy statute in *Bowers v. Hardwick* for maintaining "a lofty distance from the facts of Michael Hardwick's situation.... The distancing language [of the majority opinion]...expresses a refusal to think about how Hardwick's personhood is affected by the issue" and bespeaks "a marked effort to keep the human story at a distance" (113).

One will have one's disagreements with a book that is as imaginative, compressed, and engaged in controversial public matters as *Poetic Justice* is. There is perhaps a strand of conceptual voluntarism and bureaucratic administrativism—at odds with the book's main thrust—in the claim that novelistic thinking is "an important part of the *equipment* of a judge" (99). This seems to represent the judge as having a tool kit to be used on problems in which the judge is not directly entangled and will not be taking a distinctively political stand, in a broad sense of "political," in judgment. While this is true in many cases—for example, in many property disputes—it is less clearly true when the judge is an agent who is helping to establish the meaning of equal protection under the law in her own society.

Second, the ontology of persons in *Poetic Justice* is underdeveloped, naturally enough given its praiseworthy ambition to be a piece of public philosophy. Still, one wants to know how to balance the more or less Aristotelian emphasis on persons as like pursuers of well-being with the more or less Kantian thought that persons are distinctively agents who seek to construct their lives meaningfully in different ways, within a framework of social freedom. There is less engagement than there might be with the various strands of liberal, post-Kantian jurisprudential theory which are to be found in the writings of John Rawls, Ronald Dworkin, and Joseph Raz.

Despite Nussbaum's careful acknowledgment that other media of art might also help to cultivate valuable metaphorical imagination and that she is focusing on the novel only as an exemplary vehicle of such cultivation (6), one may wonder whether her account of metaphorical imagination is too narrow, and too careful. The novels that she treats—E. M. Forster's *Maurice* and Richard Wright's *Native Son*, besides *Hard Times*—are, while deep and significant, also conventionally realist novels that are fairly far from the more protean imaginative and linguistic efforts of William Faulkner or Gabriel Garcia Marquez or Toni Morrison. One may even wonder whether the focus on realist works is associated with a strand of administrativism and Aristotelianism more than Kantianism, in urging the thought that serious but calm reform from above, rather than revolutionary upheaval unleashing new powers of individuality from below, is what is needed and available in social life. Perhaps so; but would one want, in art or in social thinking, to do without wilder and more wayward images of suffering and redemption?

Despite these reservations, I admire Nussbaum's call for the "more consistent and humane cultivation" (xviii) of metaphorical imagination for both public and private life through the reading of densely perceptive imaginative literature. This appeal strikes me as public philosophy that is worth the name.

RICHARD ELDRIDGE

Swarthmore College

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 9, SEPTEMBER 1997

LYING, DECEIVING, OR FALSELY IMPLICATING*

A powerful tradition in ethics maintains that lying is a significantly worse choice than other forms of deception. Both choices aim for the victim to believe falsely, but only lying does so through asserting what one believes false.[1]

Deception generally, of course, need not be intentional or voluntary, as lying must. Even when it is, though, deception does not violate a simple, fundamental moral rule or principle, as does lying. Intentional deception is a constituent of many acceptable forms of everyday social life, such as tact, politeness, excuses, reticence, avoidance, or evasion, which are ways to protect privacy, promote social harmony, and encourage interest.[2] We treat putting on make-up or wearing a wig as ethically unproblematic, even when they are intentional deceptions. (For brevity, I shall write 'deception' when "intentional deception other than by lies" is meant.)

Challenging problems for the opening thesis begin to emerge, however, when we apply it to communicative deceptions that are not strictly lies. For they essentially involve, as the make-up or wig exam-

---

* Thanks to John Deigh, Georges Rey, Peter Unger, and especially Christopher Gowans for their comments, as well as to Jerrold Katz, Sidney Morgenbesser, and Frederic Schick for discussion on particular points. I am also grateful for comments by members of the audiences at the University of Miami and the Graduate Center/City University of New York.

[1] For clarification on the issues, see Frederick A. Siegler, "Lying," *American Philosophical Quarterly*, III (1966): 128-36.

[2] See Annette Baier, "Why Honesty Is a Hard Virtue," in Owen Flanagan and Amelie Oksenberg Rorty, eds., *Identity, Character, and Morality* (Cambridge: MIT, 1990), pp. 259-81. See also Immanuel Kant's "Casuistical Questions," in *The Doctrine of Virtue*, in *The Metaphysics of Morals*, Mary Gregor, trans. (New York: Cambridge, 1991), chapter II, section I.

0022-362X/97/9409/435–52

ples do not, an *invitation* by the deceiver to accept as true his deceptive contribution.

Here is an example: Abraham, venturing into a dangerous land and fearing for his life if Sarah is taken as his wife, tells Abimelech the king that she is his sister. God appears to Abimelech to warn him away from taking Sarah because "She is a married woman." Frightened, Abimelech confronts Abraham, who defends his obvious deception by denying that he lied:

> ...they will kill me for the sake of my wife. She is in fact my sister, she is my father's daughter though not by the same mother; and she became my wife....[3]

Augustine[4] defends Abraham as not having lied, classifying this as only a case in which Abraham "concealed something of the truth." Augustine does not even comment on the deception. Why do Abraham, Augustine, and many others take the lie to be so much worse, and are they correct?

In section I, aside from brief background, I offer a strategy to generalize the maneuver illustrated by this opening example. If the strategy succeeds at providing a handy way to avoid lying for deception, would it constitute, as the tradition implies, moral progress? In sections II and III, I approach this question by asking why anyone would opt to lie rather than intentionally to deceive, if the opening thesis is correct. Section II concentrates on differences in availability or assurance, and their implications for moral personality. Section III looks to broader negative social consequences. The findings tell against the opening thesis, though not uniformly. In sections IV-VI, I turn to the alleged moral distinction between lying and deception itself; and in section IV, I argue that two promising ways of drawing the lies-deception moral distinction fail in their intended application to communicative deception. There remains a felt moral difference, however, between the Abraham case and an imaginable one where instead he had lied. In sections V and VI, I defend this felt difference as a genuine ethical distinction, grounded in norms of communicative practice. But these norms can be exploited and, when they are, the moral asymmetry favoring deception is overridden.

---

[3] *Genesis* 20: 11-13 in *The New English Bible* (New York: Cambridge, 1971). See as well *Genesis* 12: 9-20.

[4] "Against Lying," in *Treatises on Various Subjects*, R.J. Deferrari, ed., *Fathers of the Church, Volume XVI* (New York: Fathers of the Church, 1952), pp. 125-79, here pp. 151-52.

The scope of this paper is restricted to the choice of deception *rather than* lying (or conversely), not the nature of either of these generally. How significant is the difference between asserting what one believes false and asserting what one believes true, inviting the drawing of a conclusion that one believes false?

I

In a wide range of cases, particularly those which pose tough problems for the view that lying is always or nearly always wrong, a (pragmatics) strategy is available to allow one to deceive rather than to lie.[5] If so, in those cases, you need not lie.

We concentrate on questions of the form 'Where is *X*?' as in the infamous case of the murderer at the door.[6] He wants to know where Joe is. You know that he intends to kill Joe, and that Joe is in the basement. You need an assertion that genuinely misleads as to Joe's whereabouts, which the murderer will accept. (Exploiting disjunctions or vague terms will generally fail on one of these criteria.)

Assume it is clearly after 11:00 a.m., though close to it. Imagine a plausible place to tell the murderer that Joe can be found, even though erroneous. If the murderer accepts it, he believes himself easily able to predict Joe's continued location, until he may intercept him—for example, being at the local bank. So you assert:

(1) If it is before 11:00 a.m., Joe is at the bank.

Since you believe the left-hand side to be false, and you accept the contentious, but credible, treatment of the truth conditions of indicative conditionals as that of the corresponding material conditional, you believe the statement to be true. So you have not lied. The murderer infers that though it is after 11:00 a.m., it is close enough so that Joe will still be there or in the vicinity, and so off he goes.

Arguably, however, in (1), as with the Abraham example, speakers do not *conversationally implicate* the conclusion drawn by the hearer. Rather, the conclusion is closer to what is asserted (or said).[7]

Here is an alternative. Select any place, well away from the basement, but fairly nearby, where Joe spends a good deal of time—for example, the local diner, the Nevada—then assert:

---

[5] Based on the work of H.P. Grice, *Studies in the Way of Words* (Cambridge: Harvard, 1989), essays 2-4.

[6] See Kant, "On a Supposed Right to Lie from Altruistic Motives," in *Critique of Practical Reason and Other Writings in Moral Philosophy*, L.W. Beck, trans. and ed. (Chicago: University Press, 1949), pp. 346-50.

[7] See Siegler, pp. 133-35; and the contrasting dissent from Grice's said/implicated distinction of François Recanti, "The Pragmatics of What Is Said," in Steven Davis, ed., *Pragmatics: A Reader* (New York: Oxford, 1991), pp. 97-120; and Kent Bach, "Conversational Implicature," *Mind and Language*, IX (1994): 124-61.

(2) He's been hanging around the Nevada a lot.

The expectation is that since the murderer believes that you are being cooperative, your answer is meant to be relevant to his question. So he will take you as having conversationally implicated that Joe is at the Nevada now.

Case (2) differs from (1) in clearly implicating only what one believes false. The risk in this strategy, however, is less certainty that the murderer will be misled. (Even for (1), if the murderer becomes suspicious, he can foil you with the question, "What makes you think that, if it is before 11:00 a.m., Joe is at the bank?")

The responses (1) and (2) are independent. It is not as if one must learn complex rules of selection. The factual information each requires is minimal and usually evident. Still, a little effort is needed to learn of these strategies. But were the alternative of lying as much worse as is alleged, we probably have an obligation to practice formulating such devious assertions.[8]

Other constructions designed to exploit what is implicated rather than stated or strictly entailed for similar ends are possible, and for other types of questions. But no claim is made that there are always feasible constructions. Also, this pragmatics strategy for deceiving cannot rival lying in ease and prospects for success. In fact, in section II, I take it as a telling difference between lying and deceiving that deceiving is more effortful. The claim is only that for answering a range of simple inquiries, this kind of linguistic deception is a workable alternative to lying.

But if these examples work in successfully misleading the murderer without lying, many will respond "So what?" At best, what I have offered is a gimmick of little, if any, ethical significance. Ordinarily understood, the intention is the same, and the means and consequences appear not terribly different.

In the passage from Immanuel Kant usually cited as capturing his view of lying within the framework of the *Grounding for the Metaphysics of Morals*,[9] he says of the false promise under the "rational beings as ends" formulation:

---

[8] Another construction akin to (1): when the basement where Joe is hiding is, say, ten miles east of the town, assert:

Joe is two miles east of the town.

Since being ten miles entails being two miles east, you have told the truth. But the murderer infers that Joe is exactly two miles east of the town.

[9] J.W. Ellington, trans. (Indianapolis: Hackett, 1981).

> For the man whom I want to use for my own purposes by such a promise cannot possibly concur with my way of acting toward him and hence cannot himself hold the end of his action (*ibid.*, Ak. 430).

The judgment applies to deception as well. Consent cannot be given, since the victim is deceived.

Still, the prohibition on asserting what one believes to be false is very strong, and on some views absolute. The prohibition on (even communicative) deception is much weaker. It is to be recommended as an admirable alternative to lying.[10] The recommendation extends well beyond philosophers, as suggested by the opening example. Political discourse is rife with misleading implications and euphemism, as with such Orwellian examples as "pacification" for mass destruction.[11]

The desperate means to avoid lying, while going easy on deception is not accidental. For if it is never right to cooperate with evil and lying is a fundamental wrong, then only a strong distinction (where deception is not seriously bad at all) will help. The hardly liveable alternative is silence in a confrontation with evil or even lesser forms of ill intent.

II

If (or when) the above strategy works and the opening thesis is correct, why would anyone choose to lie rather than to deceive? One reason is the greater assurance of success with the lie;[12] another is belief that one's untruthfulness is justified.

Even if one does quickly come up with a misleading, but truthful, assertion, to take the first reason, we noted that there is greater uncertainty of correct inferences for implicatures than for lying. But there are other differences in assurance. It is a lot easier just to lie. Unless we renounce communication, there are few effective ways to immunize ourselves against the liar. The means to lie are readily at hand and the opportunities coeval with normal conversational assertion.

Consider why trickery (or fooling) is much more thorough deception than lying. In deceiving at all, there is a violation of a norm of truthfulness. So there has to be some excuse for it, such as a good, harmless joke. But in order for it to be a joke, there has to be some display of cleverness or ingenuity. Lying leaves little opportunity to display cleverness, except as a tease. It is a blunt instrument, easily found, promising an easy success. All it takes is a suitable (contextually believable) assertion incompatible with one's relevant belief.[13]

---

[10] See the second of Alasdair MacIntyre's lucid lectures, "Truthfulness, Lies, and Moral Philosophers," Grethe B. Peterson, ed., *The Tanner Lectures, Volume XVI* (Salt Lake City: Utah UP, 1995), pp. 309-61.

[11] George Orwell, "Politics and the English Language," in his *A Collection of Essays* (New York: Doubleday, 1954), pp. 162-77, here p. 173. For many other (types of) examples, see John Wilson, *Politically Speaking* (Cambridge: Blackwell, 1990). See also *The New York Times* (September 27, 1995), B3.

[12] As elsewhere, the claim assumes contexts in which no special beliefs hold as, for example, when a speaker believes that the audience does not think he will be truthful.

But the deceiver must keep his methods fresh. Consider the device proposed here (and illustrated by (1) and (2)). Once it becomes well known or it is even suspected, it will no longer be serviceable. In legal and quasilegal settings, the option of deception is virtually eliminated by the all purpose end question: To the best of your knowledge, are there any other relevant facts?

Choice of the lie, to move to the second advantage, can reflect the liar's belief that he is justified in his act.[14] He has no need to take a devious, more effortful route, beyond what is requisite for success. The belief that one's lying is justified allows the liar a kind of honesty about his lies that is barred from the deceiver (without renunciation).

The deceiver takes a more circuitous route to his success, where lying is an easier and more certain way to mislead. The deceiver's goal is to avoid blame; so he would not forthrightly want to defend his action publicly. He takes care lest he, and not just his deception be exposed.

The "honest" liar can be open toward others as to his act with the expectation of their endorsement.[15] But in her important article "The Right to Lie," Christine M. Korsgaard[16] effectively relegates the issue to a footnote, there to be dismissed:

> Sometimes it is objected that someone could assent to being lied to in advance of the actual occasion of the lie, and that in such a case the deception might still succeed. One can therefore agree to be deceived (*ibid.*, p. 333).

---

[13] As Tyler Burge cryptically remarks: "lying for the fun of it is a form of craziness"—"Content Preservation," *Philosophical Review*, CII (1993): 457-88, here p. 474. So, too, though not as strongly, for linguistic deception. But Burge's insight overshoots if denying that sane persons can maliciously or, as a tease, lie for the fun of it.

[14] On justifying lies, see Sissela Bok, *Lying* (New York: Vintage, 1979), chapter VII.

[15] This possibility of a kind of openness about one's lying is helpful in Augustine's case of beneficent double-cross treated by Roderick M. Chisholm and Thomas D. Feehan—"The Intent to Deceive," this JOURNAL, LXXIV, 3 (March 1977): 143-59, here p. 154. Their answer will not do because it depends upon an unacceptable construal of assertion. See note 25 below.

What separates this case from standard lies is that the speaker would be quite willing to allow the hearer to view openly his pertinent highest-order intention (the intention or purpose behind the intent to represent falsely what he believes); for his highest-order intention is for the hearer's belief to be correct.

[16] "The Right to Lie: Kant on Dealing with Evil," *Philosophy and Public Affairs*, XV (1986): 325-49.

If we apply this claim to a doctor-patient relation, and if I do

> ...ask the doctor whether I am fatally ill, I cannot be certain whether she will answer me truthfully. Perhaps what's being envisioned is that I simply agree to be lied to, but not about anything in particular. Will I then trust the person with whom I have made this odd agreement (*ibid.*, p. 333)?

Contrary to Korsgaard, the answer to her rhetorically intended question, when applied realistically, is sometimes "yes." Even when the answer is "no," only full trust is eliminated. Sufficient trust may remain to sustain many social relations dependent upon it.

Social norms restrain criticism and favor complements; so we cannot be sure if a kind word is meant. But the trust is not absent, communication not dissolved. Deceptive social-science experiments proceed, even where experimenters have openly defended their studies.[17]

When practices are hardly avoidable, we accommodate. We draw rough distinctions among groups or reference classes as to whom to trust or how much, as well as when (for example, less so in matters of self-interest or privacy). Normal human psychology allows for selective attention and forgetting. We set our watches ahead, and ask others to mislead us as to the time. Certain accounts of character planning involve the unblinkered attempt to induce beliefs exploiting these shortcomings and abilities.[18]

Sometimes when one should not expect the truth, one does not know that one should not expect it. In these cases, the answer to Korsgaard's question is "yes."[19] Thus, a casual question to someone about how their family is faring can be intrusive when, unbeknownst to the questioner, there are serious marital difficulties. The trust and expectation of truthfulness are still present as the questioner does not believe that he is being intrusive, even when he endorses lies in such cases to protect privacy.[20]

But where the lie is not selected as an expression of one's belief that it is justified, lying reveals a more dangerous moral character—thus favoring the opening thesis. The deceiver, by his very effort to avoid the charge of lying, shows himself more respectful of morality, whether through honest respect or fear. The liar is more brazen. He too, of course, does not want to get caught. But what moves him is what will work best, not having the immorality attached to him after.

---

[17] The case may be thought a counterexample to my contrasts between the liar and the deceiver. Many deceptive social-science experiments are lies, however. Additionally, the willingness to engage in public justification is partly a response to external pressures. On this topic, see Bok, chapter XIII.

[18] See Jon Elster, *Sour Grapes* (New York: Cambridge, 1983), chapter IV.

[19] Compare to Korsgaard's (*op. cit.*, pp. 328-34) own clever account of why lying to the murderer can circumvent violating the first formulation of the categorical imperative.

[20] Similarly, placebos are still used in biomedical research, despite the requirement of prior consent.

The deceiver can be made to feel shame or guilt but not, or much less so, the liar. The pains that we presume Abraham suffers to find a way to avoid lying shows his extreme susceptibility to shame or guilt for moral transgression.

III

Even with this difference in character, however, the deceiver may be a greater social and moral threat than the liar. For the liar can be definitively exposed, and so, as with David Hume's[21] promise breaker, marked to the community. But thereby the deceiver is the more insidious threat (Iago being a paradigm). When exposed, he is not neatly categorized as a transgressor. After all, he never spoke falsely. He retains "plausible deniability" when he does not avoid blame altogether. Thus, "lying with statistics"[22] is attractive just because, aside from the persuasive force of precise, numerical values as the vehicle, the lying statistics are all true. That is, these are not (strictly!) lies, and that, of course, is their attraction as a way to circumvent blame.

These differences between the upfront path of the liar and the subterfuge of the deceiver are reflected in language. The liar lies. He is dishonest and not to be trusted. But the standard and slang terms for the deceiver and his means go well beyond these: sleazy, misleading, sneaky, underhanded, greasy, cunning, euphemistic, manipulative, sly, seductive, distorting, slanderous, evasive, deceitful.

Accordingly, victims do not experience lies and deception similarly. The victim of the liar, akin to the victim of coercion, has the anger and fear of one brutalized. But the brutality is clearly an external force imposed on him.

It is otherwise for the victim of the deceiver. Depending on the nature of the deception, the victim feels anything from foolish or tricked to corroded. Not only has he been misled, but the embarrassment or horror of it is that he has been duped into collaborating on his own harm. Afterward, he cannot secure the relief of wholly locating blame externally. (Consider Othello, at play's end.)

The hearer, in not demanding clarification can be held to accept the speaker's devious means, since communicational deception is overtly inference dependent. (The victim of a lie lacks even this minimal opportunity.) Of course, the deceiver is prudently betting that the hearer will not request clarification of the speaker's contribution. So, subsequently, should the hearer discover the deceit, he views himself as somewhat playing a role in it, even while his response to the assertion is natural.

---

[21] *A Treatise of Human Nature*, L. A. Selby-Bigge, ed. (New York: Oxford, 1975), book III, part II, section V.

[22] See Darrell Huff, *How to Lie with Statistics* (New York: Norton, 1954).

To the victim, the deceiver's denial that he ever lied can have the familiarly hollow ring of "Well, I never promised" or "Okay, but I didn't give my word." These are excuses to mitigate blame by exploiting a lack of explicit agreement, even when there is a failure to live up to reasonable expectations.

Much of ethics depends upon commitments that are not explicit. We care very much about the spirit of our largely tacit moral code or norms.[23] You are new to the community, and we are going together to a social function, but I do not inform you that you are dressed too casually. What powerfully, though often imperceptibly, pulls me ethically to inform you is no promise or contract or giving of my word or right, but the vast range of "our ordinary, nonexplicit commitments."[24]

IV

Two promising rationales for a lies-deception moral distinction are explicitly meant to apply to communicational deception. First, Thomas D. Feehan and Roderick M. Chisholm ask: "Why is lying thought to be *worse*, other things being equal, than other types of intended deception" (*op. cit.*, p. 153). Their answer is that since only lying is through assertion: "Lying, unlike the other types of intended deception, is essentially a breach of faith" (*op. cit.*, p. 153).[25] Second, Alasdair MacIntyre offers as the Kantian view:

> my duty is to assert only what is true and the mistaken inferences which others may draw from what I say or what I do are, in some cases at least, not my responsibility, but theirs (*op. cit.*, p. 337).

---

[23] Thomas W. Pogge's argument—in "Loopholes in Moralities," this JOURNAL, LXXXIX, 2 (February 1992): 79-98—assumes a largely explicit moral code.

[24] Stanley Cavell, *The Claim of Reason* (New York: Oxford, 1979), p. 298.

[25] Their account of assertion (*op. cit.*, p. 152), however, wrongly makes it sensitive to a speaker's assessment of the beliefs and assumptions of his audience. First, we do not accept the condition. If a person will not believe my sincere assertions, I may teach him a lesson by honestly asserting what I believe, knowing that he will not believe me. Second, the condition is unworkable. It would require speakers finely to adjust their contributions to their audience's beliefs; and it would require audiences to factor into their understanding of an assertion what they believe the speaker believes about their beliefs. A requirement for such evaluation would burden conversation with enormously complex calculational problems.

In Kant's *Lectures on Ethics*,[16] he presents an example of a person purposely packing his suitcase in front of another in order to mislead the latter to believe mistakenly that the packer is taking a trip. Had the person lied, he would be wholly responsible for the assertion and so directly for the false belief of the hearer. But the mistaken conclusions that the packer is leaving are all drawn by the observer.

The underlying idea is, presumably, that each individual is a rational, autonomous being and so fully responsible for the inferences he draws, just as he is for his acts. It is deception, but not lies, that requires mistaken inferences and so the hearer's responsibility.[17] (A plausible qualification, however, is that, if the intended victim is doing no wrong, the speaker does have an imperfect duty of assistance to correct the mistaken inferences he might draw.)

If linguistic communication involves the recovery of speaker's relevant intentions,[18] then neither of the accounts under consideration succeeds as stated. Speaker's meaning, broadly understood, is the object of the communicative intent. It is what speakers invite hearers to understand (by virtue of recognizing that very intent).[19]

The presence of an invitation crucially marks off lies from noncommunicative, though still intentional, deception. The packer does not invite the onlooker to draw any inference, nor even to understand what he is doing. He only intentionally supplies the onlooker with evidence to draw an inference. (In ignoring the expression of communicative intent or the invitation factor, Kant's dichotomy between a hearer's having a right to truth, and, his preferred, "Man has a right to his own truthfulness,"[20] is too simple.)

But this difference distinguishes not only lying from noncommunicative deception, but also, and critically, communicative from noncommunicative deception. For communicative deception, too, involves an invitation. It is the invitation to grasp what the speaker means (and not just his assertion). Since implicatures are required for the speaker not to violate the maxims of conversation, they are part of speaker's meaning (what the speaker means in saying what is said). Consequently, the invitation or communicative intent extends to implicatures.

---

[16] L. Infield, trans. (New York: Harper, 1963), pp. 224-35. MacIntyre disfavors this work as a representative Kantian text, however.

[17] The view of communication presupposed is one that sharply contrasts the decoding of assertion with all other recovered meaning as inferred. Such a view appears to guide Augustine's frequent appeals to uses of figurative language to defend Biblical characters against the charge of lying. His claim is that critics err by taking figurative language literally—"Against Lying," as well as "Lying," in *op. cit.*, pp. 45-110. But lying is possible when properly interpreting figurative language (as figurative). Teaching a child history, I assert:

Stalin was a pussycat.

Understood as intended, it is a lie.

[18] See Grice, essays 5, 6, 14, and 18.

[19] Hearers may, without error, still fail to recover what the speaker actually meant, if the speaker either inadvertently implied what he does not intend, or, more simply, genuinely violated the cooperative principle.

[20] "On a Supposed Right to Lie from Altruistic Motives," p. 347.

In taking implicatures as within the expression of a communicative intent, I have relied on H. P. Grice's own schema for working out conversational implicatures. Implicatures are posited as *required* for the speaker's assertive act to be consistent with the "cooperative principle" and the maxims falling under it (*op. cit.*, pp. 30-31).

It is unfortunately common, however, to group implicatures together with other inferences from conversational contributions. Chisholm and Feehan write of what an utterance "suggests ('implicates' or 'contextually implies'...)" (*op. cit.*, p. 155). Many inferences are not required (but optional) to avoid violation. Those are cases in which we could say that the implication is suggested, allowed, or risked. The implication is not mandated, since the assumptions violated are not *contextually mutual* to speaker and hearer, even if widely held.[31] (A test is that cancellation without reinterpretation of the assertion generates at least some puzzlement only for genuine implicatures.)

Conversational implicatures (for example, (1) and (2)) are encompassed in the speaker's invitation to his audience. Unless hearers draw these inferences, they cannot recover the speaker's relevant intentions or meaning. Appeals to autonomy and responsibility, or breach of faith, which work well to draw the line between lies and intentional deception, are then insufficient for a full moral distinction embracing communicative deception (so far as it falls within speaker's meaning). For understanding assertions requires inference, speaker's meaning includes implicatures, and misleading assertions breach the faith conveyed by a speaker's invitation.

V

We should not rest content with the results of section IV. First, in cases of (linguistic) deception, we understand the deceiver's denial that he lied as an ethical defense, even if there is something of a cheat about it. Second, while the purpose behind linguistic deception crucial to much of our discussion is to evade blame, we suppose in the case of Abraham that a purpose is to preserve moral purity. Third, sections II and III articulated systematic ethical differences between lying and deceiving. Fourth, the criticisms in section IV block only one way of finding a wrong attaching to untruthful assertion, though not untruthful speaker's meaning.[32] Specifically, as we have noted already, the lie is generally much more certain to convey intended meaning.

---

[31] See David Lewis, *Convention* (Cambridge: Harvard, 1969), section II.1; Stephen R. Schiffer, *Meaning* (New York: Oxford, 1972), section II.2; and Dan Sperber and Deirdre Wilson, *Relevance* (Cambridge: Harvard, 1986), section 1.3. In many cases, however, some subset of what is suggested is required for conformity to the cooperative principle or Sperber and Wilson's "principle of relevance."

This fourth reason needs cautious handling. That a lie is easier to ascribe than an implicated deception may give rise to a worthwhile belief in a moral distinction between them. Nevertheless, it does not establish that there is a moral distinction. We do not want to confuse the difficulty of establishing responsibility with not actually being responsible. We do not want to confuse the difficulty of showing that Abraham intended Abimilech to believe that Sarah and he were not married with his not having that intention. Similarly, we do not want to conclude from Abraham's aim to preserve purity that he avoids moral stain, rather than the avoidance being mere taboo.[33]

We need to capture and reconcile a dual reaction. From one angle, there is no moral difference. If you are going to mislead, just go ahead and lie. From another angle, the deceiver does manage to avoid a far worse wrong, even if his means are tainted.

Assistance is afforded by analogy to the distinction between doing and allowing (and related distinctions),[34] examples of which elicit a parallel duality of judgment; but the assistance is not straightforward. In falsely implicating, rather than lying, the outcome is still directly intended, not merely a foreseen consequence. Additionally, the deceiver performs an act that originates the deception, rather than merely allowing it to befall the victim.

Still, suggestive analogies do remain with the distinction between doing and allowing both in the dual reaction and in how to characterize the underlying difference. So I expect that the defense offered below of the lies-deception distinction is applicable to it; but no such application is attempted. The defense I offer pivots on specifics of communicational deception—the existence of a practice of assertion, governed by norms of truthfulness.

---

[32] Burge (*op. cit.*) forcefully argues for a transpersonal reason against lying. Lying is a disunity in reason. My impression, however, is that his argument applies to presentations-*of-meaning*-as-true, not just presentations-as-true.

[33] A concern raised by John Deigh.

[34] For a classic presentation, see Philippa Foot, "The Problem of Abortion and the Doctrine of Double Effect," in her *Virtues and Vices and Other Essays* (Berkeley: California UP, 1978), pp. 19-32.

The defense begins with the role of implicatures in tact. Implicatures have already been observed to be ways of being polite, which connects them with broader social concerns.[25] While drawing upon that connection, however, my discussion turns on a contrast between tact and politeness. The contrast is that tact is a deviation from truthfulness, requiring continuing governance by it. Consequently, it cannot seek, and would be harmed by, recognition of one's act under the description of it as tactful. But it is of the nature of politeness to want itself so recognized, as a way of promoting politeness (though it has other aims, of course). Although we want to promote tactfulness, our being tactful on an occasion succeeds only if our act is not so recognized. (It is this contrast that is valuable, not whether it perfectly discriminates tact from politeness.)

A social good promoted by tact is privacy. To develop the example suggested in section II, Larry asks Mark in a social setting: "Where's Laura?" (Laura is Mark's wife.) Mark answers: "Oh, she is away on business." In fact, Laura and Mark have just legally separated, but Laura happens to be away on a business trip. Mark did not lie. Nor did he reject the question, which would be socially unpleasant, as well as self-defeatingly hinting at marital difficulties.

The inference that everything is all right between Laura and Mark is immediately drawn by Larry. If subsequently, Larry discovers the truth, there is latitude for him to attribute Mark's misleading comment to reasonable distraction. But even in the absence of the excuse, Larry recognizes that Mark did not lie, since Laura is actually out of town. Given the bind that Mark was in, his answer is appropriately tactful, and even admirable if an effortful attempt to avoid an outright lie. So the opportunity to deceive smooths over social frictions without the naked transgression of a lie.[26]

The contrast between tact and politeness applies. It is retrospectively, not during the act, that the tactfulness is so recognized. Even where we anticipate a tactful environment, as noted in criticizing Korsgaard, truthfulness is still a governing, if attenuated, norm. So the smooth functioning of conversational practice is not obstructed. Truthfulness is still governing communication, wherein tactful comments are only a minor fragment.[27]

---

[25] See Mark Kingwell, "Is It Rational to Be Polite?" this JOURNAL, XC, 8 (August 1993): 387-404.

[26] See also Sperber and Wilson, p. 62.

[27] See Lewis, "Languages and Language," in his *Philosophical Papers, Volume I* (New York: Oxford, 1983), pp. 163-88.

Continued governance by the norm of truthfulness forcefully reveals itself as a limit on tact (and implicatures). We can ask a pointed question, insisting on an explicit, candid answer.[38] Since the concern for true information is so fundamental, it is infuriating when a request for honest opinion is replied to with what one regards as tactful evasion or flattery. (Unlike Grice's other maxims, ineliminable violation of truthfulness violates entitlement to assertion.)

The value of tact, as revealed in the previous example, goes beyond smoothing social frictions, allowing one "put on the spot" to escape gracefully. It lessens the "strains of commitment,"[39] due to the hardly avoidable social situations that encourage lying.[40] So in allowing for tact, we sharpen the contrast with falsely asserting. The harsh demands upon us to preserve our honesty are diminished, and so compliance with the rule against lying is promoted. We do not have to choose between being socially brutish or sacrificing our privacy or just lying.

Alternatively, one could extend a proposal of Henry Sidgwick's:[41]

> ...it should be universally understood and expected that those who ask questions which they have no right to ask will have lies told to them (*ibid.*, p. 318).

But tactfulness is preferable to extending Sidgwick's blunt response either to cases of innocent intrusions or to replace social graces. Better to compromise on truthfulness than to abandon it or be oppressed by it, especially when the compromise sharpens the contrast with lying, thereby reinforcing the prohibition on it.

VI

Encouragement to tact and politeness depend upon tacit (and fragile) social arrangements, however common. But let us now concentrate on primarily informational aims of communication. Implicatures predominantly function to facilitate economical conversation (saying more with less).[42] Consider the simple exchange:

---

[38] Thomas Hill, Jr. observes that evasion and related linguistic deceptions afford the opportunity for more precise questioning, while lies do not—"Autonomy and Benevolent Lies," in his *Autonomy and Self-Respect* (New York: Cambridge, 1991), pp. 25-42, here p. 41. But this opportunity applies only when a listener becomes uncommonly aware.

[39] John Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971), sections 25 and 29.

[40] One reason Pogge's criticisms (*op. cit.*) do not apply to the loophole of false implication is the ordinariness (in harms) of the demands that the loophole mitigates. Also, the moral systems permitting the loopholes defended are overall beneficial, while Pogge denies the truth or force of this suggestion for the cases that he discusses.

[41] *The Methods of Ethics* (Indianapolis: Hackett, 1981, 7th ed.).

[42] Another difference is that tact is partly taught and a late social development. Use of implicatures stems from rudiments of language learning that begin, in fact, prelinguistically. For prescient anticipation of "developmental pragmatics," see Stanley Cavell's "Must We Mean What We Say?" in his *Must We Mean What We Say?* (New York: Scribner's, 1969), pp. 3-43.

Ted: "Whose turn is it to walk the dog?"
Marcia: "I did it yesterday."

In this brief response, Marcia implicates both her answer (it is Ted's turn) and her justification (the fairness of taking turns).

Dan Sperber and Deirdre Wilson contend that these implicatures being obviously intended are just as much Marcia's responsibility as if she had asserted them:

> ...the speaker is committed to the truth of all determinate implicatures conveyed by her utterance, just as much as if she had expressed them directly.*

Underlying this claim is the assumption that responsibility is solely proportioned to determinacy of the implicature. This reduction of ethical to agent-causal responsibility looks to only one side of our dual response to deception, the side of the individual act and actors, not the practice.

Dissent from Sperber and Wilson's view is off center, however. Implicatures arise under diminished *demands* of truthfulness, not expectations. Hearers are less disposed to challenge or request reasons. But the expectation of truthfulness remains; for were it to diminish, so would informational value. The lessened demands of truthfulness being mutually expected, rational, and tacitly endorsed have normative weight.

In the example just offered, Marcia (politely) implicates a turn-taking rationale, and so licenses Ted to infer that it is his turn. In example (2), what you assert ("He's been hanging around the Nevada a lot") supplies the ground for your answer, and so licenses the hearer to draw the direct answer to his question. Speakers supply their reasons, licensing hearers to draw the conclusion from it. So in a single economical response, speakers yield both their reason and the relevant information.

But we also lessen the demands of truthfulness, even if the implicature is determinate. First, since the speaker offers his reason for the purpose of drawing a conclusion, it is that conclusion, not its ground, that is the focus of attention. Second, since the hearer draws the inference—detaches the answer to his question—he endorses it to that extent. Consequently, if the hearer comes to question the reason, he questions his own endorsement; and if the speaker's purpose is to mislead, the hearer admits his own complicity.

---

* "Inference and Implicature," in Davis, *op. cit.*, pp. 377-92, here p. 384; see also p. 387, and *Relevance*, p. 199; and Grice, pp. 39-40.

These two features hold widely of implicating. Implicatures are overtly dependent upon hearer-drawn inferences, whose conclusion is most germane (compared to the premises) in the exchange, and so the focus of attention.

Consider Grice's example:

A: Smith doesn't seem to have a girlfriend these days.
B: He has been paying a lot of visits to New York lately (*op. cit.*, p. 32).

The implicature is that Smith [probably] has a girlfriend in New York. You, too, readily draw this implicature, while imagining yourself as a participant (*A*), not merely a professional reader.

Contrast your response if *B* had instead replied:

B: Smith [probably] has a girlfriend in New York.

Now you, as *A*, are more inclined to ask:

A: How do you know?

To which *B* offers his original response, as his focal reason. But now you (*A*) are much more disposed to express doubts as to whether there is a reliable connection between a man's visiting New York and his having a girlfriend there, assuming that there is no special (mutual) knowledge about Smith overruling these doubts. In the original dialogue, however, you have these doubts, too. Yet they are unlikely to become salient in that exchange, and even if salient, they appear less appropriate to offer in challenge.

Your (*A*'s) challenge to the assertion in the form 'How do you know?' is, as well, standard and revealing. (Or, 'Can you prove [show] that?'[44]) But to a case of implicating that *q*, where *q*'s being implicated becomes cognizant to the hearer, the typical challenge would take the form:

Do you know that *q*?

or

'Do you mean that *q*?' ('Are you saying that *q*?')

Even when it is fairly obvious (determinate) to the hearer what the speaker is recommending, only the latter form is usually pressed. Clearly, one cannot challenge a speaker for a reason when one cannot yet determine that a claim has been entered.

---

[44] These forms are much more standard than 'Do you believe that *p*?', even if the latter is favored when *p* is an opinion or controversial claim. If standard, it supports Peter Unger's hypothesis that in asserting that *p* we represent ourselves as knowing it—*Ignorance* (New York: Oxford, 1981), pp. 260-65.

Speakers often have information which they believe useful to hearers, but which they are reluctant to offer if placing themselves in a position to have to defend it fully. Think here of familiar uses of 'It may be that *p*'. Speakers intend for hearers to take seriously the possibility that *p*, while likely exempting themselves from challenge by affirming only the (mere) possibility that *p* is true. The purpose need not be to insinuate *p* without taking responsibility for it. Rather, it is that a full defense of *p* would just be too much of a burden or effort.

So a norm corresponding to the lessened demands of truthfulness for implicatures would be desirable for all. Such a norm of conversation acquires moral force. Given the previously mentioned "strains of commitment" generated by the numerous situations pressuring us to deceive, it is rational for each of us to want cooperative accommodation with others. The difference in demands of truthfulness for assertions compared to implicatures provides a salient rationale for a corresponding ethical norm. For the difference is an endorsed mutual expectation under which our contributions are offered and accepted, or occasionally challenged. (Recall the starting point as supportive evidence: we do have the attitude that the lie is worse.[43]) The norm is suitably circumscribed. Since the expectation of truthfulness remains, no simple permission is extended for deception.

If there is an ascribed difference,[44] a difference based on our endorsements and cooperative understanding, then this will account for one side of the dual reaction noted earlier. It is the reaction that the lie is worse. But we can also account for the opposed reaction as due to our searching in vain for a moral difference in features of the specific intent or act alone.

To conclude this section and the paper: in section IV, I assimilated the message *in* implicatures with assertion as both features of extending the communicative intent. But I there overlooked the message conveyed *by* the very choice, rather than assertion, of implicatures. One message we just observed is of diminished demands of truthfulness.

---

[43] In the debate over doing and allowing, our analogous attitude (favoring allowings) is not even treated as prima facie evidence of a distinction. It is explained away as resting on happenstances of actual cases, especially that doings are generally more certain of their outcome than allowings. The standard opening move is just to assume that doings and allowings are equally certain of achieving their ends. But this is to overlook the possibility that a systematic, even if incomplete, epistemic difference can be a basis for a distinction that we ascribe.

[44] Such as a difference in *symbolic value*. See Robert Nozick, *The Nature of Rationality* (Princeton: University Press, 1993), pp. 26-35.

Analogously, the choice of a deception rather than a lie can itself convey a message of moral effort and struggle indicative of a genuine distinction. A lies-deception distinction corresponding to an assertion-implicature one is rational for us to want (for purposes of social harmony, lessening the strains of commitment, and facilitating the exchange of information). For these goals a norm is needed, since knowledge in advance, rather than on-the-spot calculation, is the only feasible way that the distinction could accomplish its goals.

The account fits the dual response observed earlier. There is a moral asymmetry favoring deception to lies. The means of misleading matters. But the account also explains the opposed reaction. If one looks just at the particular case, no inherent difference is noticed.

But the purposes for which the norm is adopted need not be the ones for which one relies upon it. Thus, the lingering worry that the deceptive speaker exploited an ineliminable loophole of a fair understanding can fester. Especially if the loophole becomes too readily available, as with the false implicature strategy of section I, it ceases to be a safety valve to diminish pressures to speak falsely. It becomes a means of evasion and an attraction in itself. The moral effort to avoid lying diminishes and the expectation of truthfulness is harmed.

Similarly, I concluded in section III that where our implicit commitments are a strong governing force, a defense of one's action which relies primarily on the lies-deception distinction teaches a bad moral lesson. It encourages deviousness and a legalistic attempt to get away with what one can. One may be less blameworthy for deceiving than lying, but still have acted worse by seeking to be less blameworthy. One is not attempting, as with Abraham, to act as nobly or honorably as harsh circumstances permit. Even the good person, who exploits this distinction only when his back is up against the wall will be shamed (so that one imagines Abraham, when confronted by Abimelech, as squirming). Consequently, one can acknowledge that there is a moral asymmetry favoring falsely implicating or deceiving to lying, while holding neither that this is true in each case nor that a general deceptive strategy to avoid lying is moral progress.

JONATHAN E. ADLER

Brooklyn College and the Graduate School/
City University of New York

PUTNAM'S BORN-AGAIN REALISM

Hilary Putnam's 1994 Dewey Lectures[1] take us on a quick tour of his positions on realism, moving from the earliest Putnam, who was—he informs us—a naive realist, back again to a new "natural" or common-sense realism that he dubs a "second naiveté." The first two lectures clear the way to a common-sense realism in perception, drawing on arguments by John Austin and Ludwig Wittgenstein for help in dislodging "disastrous" theories of perception that interpose a sense-datum "interface" mediating between our consciousness and the external world. Apart from stating that I generally favor it, I shall say nothing about perceptual realism, focusing instead on Putnam's cognitive realism developed in the third and final lecture. Putnam there argues that conception, like perception, is veridical: true cognitive judgments, including unverifiable judgments about the past, are given the naive realist interpretation that he now champions. Putnam again claims Wittgenstein's support: "When we say, and *mean* that such-and-such and is the case, we—and our meaning—do not stop anywhere short of the fact; but we mean *this-is-so*" (492-93).[2] In an essay on Wittgenstein, published the same year as the Dewey Lectures, Putnam "naively" interprets 'Caesar crossed the Rubicon' in the light of the doctrine that "understanding a language-game is sharing a form of life."

> I think it is undeniable that our ordinary linguistic practice in connection with such a statement is deeply informed by a realist "picture." For educated persons, it is virtually as if we had seen Caesar crossing the river.[3]

The living use of a sentence suffices to "reach" a fact, even one "distant from the speaker" and even (as against Michael Dummett) when assertability conditions are lacking, as they are for 'Lizzie Borden killed her parents with an ax':

> If we accept it...that we are able to conceive of how things that we cannot verify *were*—then it will not appear as "magical" or "mysterious" that we can understand the claim that the sentence is *true* (510-11).

---

[1] "Sense, Nonsense, and the Senses," this JOURNAL, XCI, 9 (September 1994): 445-517. Subsequent parenthetical page references in the text are to this article.

[2] Quoted from *Philosophical Investigations*, G. E. M. Anscombe, trans. (New York: MacMillan, 1953), §95.

[3] "Does the Disquotational Theory Solve All Problems?" in *Words and Life*, James Conant, ed. (Cambridge: Harvard, 1994), p. 276.

0022-362X/97/9409/453–71

Next best to having been there is to hear or to read about Caesar's crossing the Rubicon or Lizzie's axing her parents or Dwight Eisenhower's receiving the German surrender in 1945 (492).

We are invited to regard Putnam's advent to a second naiveté as the latest and perhaps final stage of a journey back to the commonsense realism in which he once believed. On the whole, realists like myself who have never left home would be disposed to welcome him. Given Putnam's well-known hostility to "metaphysical" realism and his long dalliance with irrealisms, however, we shall want assurance that the prodigal has returned to the right address. For Putnam's current realism continues to reject what he (and most others) regard as an essential element of traditional realist doctrine. He had earlier said: "Whatever else realists say, they typically say they believe in a correspondence theory of truth."[4] Yet we still find him inveighing against correspondence in the grand manner of his irrealist days, calling correspondence realism a "fantasy," or "funny," "mysterious," "metaphysical," "queer," and so forth.

On the older realist view, a judgment to the effect that a certain proposition is true says something about how the proposition is related to reality; the correspondence realist understands the judgment that *p* is true as the claim that some feature of reality—call it "the fact that *p*"—corresponds to and "makes true" that *p*. Throughout the journey that takes us back to Putnam's naive realism of cognition, he repeatedly takes pains to dissociate his own new realism from this classical correspondence realism; so whatever realists may "typically say," Putnam now holds that a proper realist must abjure correspondence. Donald Davidson,[5] too, regards classical realism as based on correspondence; and he, like Putnam, rejects it.

> The realist view of truth, if it has any content, must be based on the idea of correspondence, correspondence as applied to sentences or beliefs or utterances—entities that are propositional in character; and such correspondence cannot be made intelligible (*ibid.*, p. 304).

My aim in what follows is not so much to provide arguments for correspondence realism as to give it the kind of formulation that protects it from Putnam's (and Davidson's) animadversions. In aid of this, I shall show that Putnam, in common with most other analytic philosophers, misconstrues the nature of the ontological relatum in the traditional theory, which takes truths to be in "correspondence" with extralinguistic ontological features of reality.

---

[4] *The Many Faces of Realism* (La Salle, Ill.: Open Court, 1987), p. 31. For Putnam, a belief in correspondence is integral to being a "metaphysical realist." Thus, he says of W.V. Quine that he cannot be "a 'metaphysical realist' in my sense, since he does not accept the correspondence theory of truth."

[5] "The Structure and Content of Truth," given as Dewey Lectures in 1989, published in this JOURNAL, LXXXVII, 6 (June 1990): 279-328.

I

In the first lecture, Putman reminds us that he objects to W. V. Quine's[4] doctrine that 'exists' is univocal in such different statements as 'Brick houses on Elm Street exist' and 'Prime numbers greater than one million exist' as reflecting a view of language that is "deeply misguided" (450fn.). The third lecture says the same about 'true' (514fn.). Putnam there says that 'true' is not univocal when said of assertions about houses and numbers, for such assertions belong to different language games.

> Our understanding of what truth comes to, in any particular case (and it can come to very different things), is given by our understanding of the proposition, and that is dependent on our mastery of "the language game".... [K]nowing what truth is in a particular case depends on knowing the use of signs in the language game... (513).

Calling correspondence "a mysterious relation that somehow underwrites the very possibility of there being knowledge claims," Putnam disparages the idea of a univocal "free-standing property of truth" belonging to all true assertions in virtue of "their correspondence to reality." Davidson located the trouble on the ontological side:

> The real objection [to correspondence theories] is that such theories fail to provide entities to which truth vehicles (whether we take these to be statements, sentences, or utterances) can be said to correspond (*op. cit.*, p. 304).

The "entities" wanted are usually called facts. Analytical philosophers have found the notion of "fact" dubious and even incoherent. It must be admitted that traditional realists who had attempted to clarify how truth is related to fact had not been helpful.

To his Wittgensteinian scruples concerning a notion of truth common to such different propositions as the ones about brick houses and large prime numbers, Putnam adds a deflationism that fragments the univocity of truth in a more radical way. For he holds that in any given case "if I assert that 'it is true that *p*', then I assert the same thing as if I simply assert *p*" (501). As a realist, Putnam believes we can truly assert propositions for which there may be no warrant. The combination of anti-Dummettian realism and deflationism leads him to say that what 'true' means, if anything, differs from sentence to sentence even within the same language game.

---

[4] *Word and Object* (Cambridge, MIT, 1960), p. 131.

Assume that 'There are brick houses on Elm Street' and 'There are no cows on Park Avenue' are both true assertions in the same language game. Since by Putnam's deflationism, the truth of one "comes down to" there being brick houses on Elm Street and of the other to there being no cows on Park Avenue, the judgments that both are true come to different things. Deflationists have therefore faced the question why we need a truth predicate at all. The usual answer is that—apart from its use to emphasize or endorse—'true' provides a way of expressing disjunctions and conjunctions that could not otherwise be expressed. Hartry Field[7] gives the example of a philosopher, opposed to Dummett, who believes there are sentences that are true though we have no way of justifying them. Without a truth predicate, he would have no way of expressing his thesis; for it would require an infinite disjunction of sentences of the form 'It might be the case that *S* even though we may never have reason to believe it'. Using 'true' gives him a way: "For," says Field, "the claim that it might be the case that some sentence in our language is true but that we will never have reason to believe it can be viewed as simply a way of formulating the disjunction" (*ibid.*, p. 264).

Except for such uses, however, the truth predicate is not needed. It is not needed for any individual sentence, since in any given case, '*p*' itself suffices to say what 'It is true that *p*' says. We do nevertheless predicate 'true' of individual sentences saying of 'There are no cows on Park Avenue' that it is true and of 'There are brick houses on Elm Street' that it is true. In all such cases, 'true' means different things. For in saying that *p* is true and that *q* is true we are simply saying that *p* and saying that *q*, and in each case we say different things.

Putnam is a deflationist, but he notes disapprovingly that deflationists often share with Dummett the view that understanding a sentence consists in the ability to recognize whether it is verified. The deflationist, says Putnam, "is therefore unable to capture the sense in which certain statements about the past (namely, the true ones) are fully as *right* as statements about the present" (499). He rejects that side of deflationism.

> What is wrong in deflationism is that it cannot properly accommodate the truism that certain claims about the world are (not merely assertable

---

[7] "Deflationist Views of Meaning and Content," *Mind*, CIII (July 1994): 249–85.

> or verifiable but) *true*. What is right in deflationism is that if I assert that "it is true that *p*," then I assert the same thing as if I simply assert *p* (501).

Rejecting "what is wrong in deflationism," Putnam is a realist in the sense of 'realist' that Dummett is an antirealist. Accepting "what is right in deflationism," Putnam is an antirealist in the sense that the classical correspondence theorist is a realist. For the correspondence realists deny that predicating truth of a proposition is merely tantamount to asserting the sentence that expresses the proposition, holding that when I say it is true that there are brick houses on Elm Street and or true that there is an even prime number, I speak not of houses or of numbers but of propositions, saying of them that they correspond to reality.

> [T]he metaphysical realist needs to argue that there is something we are saying when we say of a particular claim that it is true over and above what we are saying when we simply assert the claim. He wants truth to be something that *goes beyond* the content of the claim and to be that in virtue of which the claim is true. This forces the metaphysical realist to postulate that there is some single thing we are saying (over and above what we are claiming) whenever we make a true claim... (501).

Although Putnam calls it "queer," he correctly attributes to metaphysical realists the view that predicating truth of a statement "goes beyond the content" of the statement so that any claim of the form 'It is true that *p*' is understood as a claim that '*p*' corresponds to—is made true by—some feature of reality.

G. E. Moore[8] is a good example of the kind of realist Putnam has in mind. He talks of true propositions as "true beliefs"; the facts to which they correspond he sometimes calls facts and sometimes "truths":

> Whenever I consider a true belief, no matter what the nature of the belief may be, I can't help thinking that there is always something in the Universe one of those peculiar things which I call truths, corresponding to it (*ibid.*, p. 311).

As a correspondence realist, Moore believed that diverse facts (or "truths") share a common nature that fits them for the role of truth makers and he asked: "What, after all, is the property which all truths have in common and which is not shared by anything which is not a truth" (*ibid.*, pp. 309ff.).[9] Putnam, on the other hand, rejects "the metaphysical idea that there is a 'substantive property' whose existence underwrites the very possibility of using the word 'true'" (502).

---

[8] *Some Main Problems in Philosophy* (New York: MacMillan, 1953).

[9] Moore "confesses" he does not know the answer, but adds that though it is not easy to specify their nature, "we all know perfectly well" what facts are. My answer to Moore's question is given in section III below.

As a deflationist, Putnam holds that each fact is as particular as the true statement that states it. As for correspondence, though we may choose to say (as even Wittgenstein sometimes does) that a true proposition "corresponds to a reality," saying so never takes us beyond the particular truth claim. For we are still saying no more that what was said in making the assertion. Putnam asks what we could mean by saying that an assertion like 'There are no intelligent extraterrestrials' corresponds to a reality. He notes: "One can say what reality corresponds to it, if it is true, only by using the words themselves" (503). Putnam makes the same point about Wittgenstein:

> Wittgenstein says that "This chair is blue" (imagine he had a blue chair in front of him) corresponds to a reality, but he can only say to what reality by using the sentence itself (513fn.).

It is in the nature of the language game that its true statements are factual reports. That suffices for a naive cognitive realism without appeal to an "underlying reality" to which a true statement "corresponds." As Putnam interprets him, Wittgenstein rejects "the metaphysical realist's understanding of such platitudinous thoughts as the thought that 'This chair is blue' can correspond to the fact that a particular chair is blue" (513).

Putnam concludes the Dewey Lectures by assuring us that the naive realist can safely jettison the older realist assumption of "a funny metaphysical something standing behind our talk," since "giving up on the funny metaphysical somethings does not require us to give up on concepts that, whatever our philosophical convictions, we employ and must employ when we live our lives" (517).

In renouncing correspondence realism, Putnam belongs with the majority of clear-headed philosophers since Gottlob Frege. On the other side are the few clear-headed friends of correspondence—chief among them, Austin (praised by Putnam in the first two lectures for championing a common-sense perceptual realism, but tactfully unmentioned in connection with the new cognitive realism of lecture III that Putman espouses). Austin and other correspondence realists did not succeed to arrive at a coherent conception of the ontological relatum of correspondence. I consider this unfortunate. Because the notion of truth-making facts was left in a muddle, facts soon became, in Quine's words, *entia non grata* and realism, in the twentieth century, while given lots of lip service, remained undeveloped as a coherent philosophical doctrine.

According to P. F. Strawson, Frege, Quine, Putnam, and Davidson, the thesis that a true proposition corresponds to some nonlinguistic correlate that "makes it true" is not merely unexplained but inexplicable. There is general agreement that Strawson[10] was right when, speaking of "truth-making" facts as the nonlinguistic relata of correspondence, he said: "The demand that there should be such a relatum...is logically absurd." The problem, he pointed out, was that while things like cats and mats are out there, a fact like the cat being on a mat is not.

> The only plausible candidate for what (in the world) makes a statement true is the fact it states; but the fact it states is not something in the world (*ibid.*, p. 195).

Citing Strawson with strong approval, Davidson also concludes that "the idea of correspondence...is an idea without content" (*op. cit.*, p. 305).

By contrast, the traditional realist maintains that a fact like a cat being on a mat is no less real (and no less "nonlinguistic") than the cat on the mat and this fact makes 'There is a cat on the mat' true. The correspondence realist is well aware that facts are categorially different from objects; for example, it makes no sense to speak of the weight or color of a fact. That facts are unlike objects is one reason modern philosophers have found it easy to dismiss them.[11] The more pertinent reason is that traditional realists, for whom facts play the crucial role of truth makers, had failed to specify their nature.

Because modern philosophers from Frege to Putnam have renounced truth-making facts, an acceptable sense of 'fact' that has for some time been current is as a synonym for 'true sentence', 'true statement,' 'true proposition', or 'true thought'. In any such variant, "facts" are not "truth makers" but truths that themselves stand in need of an account of what makes them true. The accounts of truth making that abjure truth-making facts are broadly "compositional" and object oriented; the truth of a sentence is a function of the denotation of its parts (Frege), or of the ways its predicates are satisfied by sequences of objects (Alfred Tarski).

---

[10] "Truth," in *Logico-Linguistic Papers* (London: Methuen, 1971).

[11] Indeed, as I shall later argue, facts are not in the world, but have reality as properties of the real world. Many things in the world (shaped, colored, textured, and so on) can be represented. We can, for example, mentally visualize or physically represent some gray cats. But the fact of their existence to which a true statement like 'some cats are gray' corresponds is not in the world and not the kind of thing that can be represented.

Davidson says that because representationalism is false, so is correspondence: "It is good to be rid of representations and with them the correspondence theory of truth"—"The Myth of the Subjective," in Michael Krausz, ed., *Relativism, Interpretation, and Confrontation* (Notre Dame: University Press, 1989), pp. 165-66. That is both wrong and the wrong way round: correspondence is true but representationalism in any straightforward sense is false because the facts to which true statements correspond cannot possibly be "represented."

II

In the ensuing discussion, 'sentence' and 'statement' will mostly be used interchangeably but they are formally distinguished. An indicative sentence assertively uttered is a statement. Statements express propositions. For example, assertively uttered, the sentence 'There are no elves' is a statement that expresses the proposition THAT THERE ARE NO ELVES. (In what follows, 'sentence' will always be understood as 'asserted sentence' or statement. The proposition it expresses will be represented in upper case.) A statement claims truth for the proposition it expresses and is itself said to be true when the proposition it expresses is true. For example, if I say 'I'm hungry' and my son David says 'My dad's hungry', both statements are true if the proposition they express in common is true.

Once expressed, the proposition exists. Saying something no one has said before is like whistling a tune never before heard or thought of. The tune does not exist prior to the act of whistling it. Similarly, no proposition exists except as the product of an initial speech act such as saying, asking, or promising.[2]

According to Rudolf Carnap, a term like 'gray' signifies a property or characteristic: being gray or grayness. According to Frege, it expresses a "sense." I prefer to say that 'gray' expresses the "characterization" BEING GRAY, but, in any case, I believe it is important to recognize "signifying" along with "expressing" as two distinct modes of meaning.

A term or sentence in use *expresses a characterization* which we represent in upper-case letters. For example, the term 'gray' expresses BEING GRAY. It *signifies a property or characteristic* (for example, being gray) and it may denote something to which the characterization expressed applies (for example, the old gray mare).

The sentence 'Some mares are gray' expresses the propositional characterization THAT SOME MARES ARE GRAY or SOME MARES BEING GRAY and it signifies that some mares are gray—a characteristic of the world. For the world would be different if there were no gray mares.

---

[2] The orthodox (Fregean) view is that propositions or thoughts are "eternal." Prominent among those holding the view that propositions are human products is Karl Popper, who regards them as "world 3" constituents. Nominalists (for example, Quine, Israel Sheffler) do not "countenance" propositions in either sense.

The characterization, BEING GRAY, corresponds to the characteristic of being gray. Because there are no properties that nothing and no one possesses, a term may express a characterization but fail to signify a corresponding characteristic. Since there are no elves, there is no such characteristic as being an elf; and the term 'elf', which expresses the characterization BEING AN ELF, is vacuous with respect to both signification and denotation. Similarly for sentences: 'There are elves' expresses the propositional characterization THAT THERE ARE ELVES, but fails to signify a corresponding characteristic of the world.

The distinction between characterizations and characteristics was carefully observed by Aristotle and by medieval philosophers, who remarked that the same verbal expressions are equivocally used to refer to both. Thus, 'being wise' is sometimes used to refer to a characterization or "predicable" (as in 'Being wise correctly characterizes Socrates') and sometimes to a characteristic or "universal" (as in 'Being wise is a characteristic of Socrates'). The term 'wise' expresses BEING WISE, which, says Peter of Spain,[13] is "suited to be said" of something. The same term signifies being wise or wisdom, a "universal" which is "naturally suited to be in" something as a property or characteristic of that thing.

> 'Predicable' taken in its strict sense and 'universal' are [vocally] the same but they differ in that a predicable is defined in terms of saying and a universal in terms of being (*ibid.*, pp. 80–81).

We find the same kind of characterization/characteristic equivocation in an expression like 'that there are flying mammals', which is sometimes used to refer to THAT THERE ARE FLYING MAMMALS (a propositional characterization) and other times to refer to the fact that there are flying mammals (a characteristic of the real world).

The classical realist distinguishes the corresponding senses of 'fact'. In one sense, a FACT is a correct characterization of reality, a true statement or true proposition.[14] In another (lower case, and as yet unexplicated) sense, a fact is a determinate characteristic of reality. What makes a given characterization true is said to "correspond" to it. According to the correspondence realist, facts "correspond" to FACTS, which is to say: facts make true propositions true. For example, there being flying mammals (a fact) makes THERE BEING FLYING MAMMALS (a FACT) true.

---

[13] "Predicables, Categories," in *The Cambridge Translations of Medieval Philosophical Texts, Volume I, Logic and Philosophy of Language* (New York: Cambridge, 1988).

[14] 'Characterization' is itself ambiguous being used for expressions and for what they express. The proposition THAT THERE ARE ELVES is a characterization but so, too, is the sentence, 'There are elves,' that expresses it. 'FACTS' and 'truths' are similarly ambiguous; 'There are elks' is said to be a FACT and so is the true proposition THAT THERE ARE ELKS. Again, we speak of an expression like 'is wise' and of the predicable BEING WISE it expresses as characterizations. Both are said to characterize things possessing the characteristic of being wise (wisdom). But because quotation is almost invariably used when referring to expressions, the ambiguities are generally harmless.

Moore, a most sophisticated naive realist, remarked on the ambiguity of phrases of the form 'that *p*', noting that

> [W]e can only name beliefs by means of expressions beginning with 'that'—that lions exist, that bears exist—and so on. But curiously enough if we want to name the fact to which a belief refers...if the belief be true,...we can only do it by means of exactly the same expressions (*op. cit.*, p. 257).

Moore accordingly spoke of the "dangerous" ambiguity "which attaches to the expression 'fact'." To disambiguate, he often referred to propositional "facts" as "true beliefs" and to the corresponding ontological facts as "truths."

Unlike Moore, Putnam fails to appreciate the systematic ambiguity of the expressions we use in talking about the two terms of the correspondence relation; because of that failure, Putnam's deflationist "same words" argument against correspondence misses the mark. Consider his comment on the statement 'There are no intelligent extraterrestrials': "One says what reality corresponds to it if it is true, only by using the words themselves." The comment portrays correspondence realists as spinning their wheels when they claim that the statement corresponds to a reality. But it begs the question. The correspondence realists hold that THERE ARE NO INTELLIGENT EXTRATERRESTRIALS, if true, is true because it corresponds to an "underlying reality" that makes it true, namely, to the (lower case) fact that there are no intelligent extraterrestrials. That "the words themselves" just used in speaking of the FACT are also conventionally used in referring to the fact is no reason to impugn their correspondence. Consider again that "the very same words" serve as subject expressions in 'BEING WISE is a (true) characterization of Socrates' and 'Being wise is a characteristic of Socrates'. That the same phrase is used is no argument for denying the truism that the predicable (BEING WISE) is true of Socrates because he possesses the characteristic of wisdom.

Putnam also uses the "same words" argument to deflate Wittgenstein's explicit talk of correspondence, noting that when he says 'This chair is blue' is true because it "corresponds to a reality...he can only say to what reality it corresponds by using the sentence itself." But here again, Putnam is inattentive to the conventionally equivocal use of phrases of the form 'that *S*' and so fails to distinguish between FACT and fact, between THAT THIS CHAIR IS BLUE (the FACT) and that this chair is blue—the truth-making fact to which the FACT "corresponds."

I have no opinion on whether Wittgenstein was any kind of realist. I am suggesting, however, that in reading Wittgenstein as one, Putnam may be getting more than he bargained for. For he dismisses Wittgenstein's overt talk of "correspondence to a reality" and assumes, without offering much by way of ponderable evidence, that Wittgenstein's realism is the kind Putnam finds acceptable: antiverificationist but also anticorrespondence. Not only does Wittgenstein occasionally speak just like an old-fashioned correspondence realist, in *On Certainty*[13] he carefully distinguishes the two classical terms of correspondence as objects of different epistemic attitudes. I may believe truly THAT THIS CHAIR IS BLUE and know that this chair is blue. What I know is a fact, but what I believe is just the proposition expressed by a sentence (what Frege called its sense):

> "I know" is supposed to express a relation, not between me and the sense of a sentence (*einem Satzsinn*) but between me and a fact (*einer Tatsache*) (*ibid.*, paragraph 90).

Wittgenstein here implicitly recognizes a "real distinction" between FACTS and facts; we may believe a FACT and know the corresponding fact, but the fact is "a reality" and we cannot logically speak of believing it, nor can we "know" a (true) proposition. Wittgenstein seems to hold that a true sentence not only expresses its sense (a FACTUAL *Satzsinn*); it also signifies a fact (*Tatsache*). If so, Wittgenstein's talk about "corresponding to a reality," which Putnam dismisses as not serious, may well be the old-fashioned correspondence realism of Moore and not the "naive" correspondence-free realism that Putnam says he "wish[es] to attribute to Wittgenstein" (499).

We have yet to say what facts are, but even now we may say that the same failure to hold fast to the distinction between a FACT (as that which a true statement states) and the fact that makes it true vitiates Strawson's remark that the "only plausible candidate for (what in the world) makes statements true is the fact it states; but the fact it states is not something in the world." Having conflated truth-making facts with the FACTS that true sentences express, Strawson concludes that correspondence theories are "logically absurd."

---

[13] G. E. M. Anscombe and G. H. von Wright, trans. (New York: Harper, 1972).

That said, we again acknowledge that classical realists never did give an acceptable account of the reference of such phrases as 'the fact that there are flying mammals' or 'the fact that there are no flying turtles'. The classical realist owes us a clear answer to the question: To what do such referring phrases refer?

It is not generally realized that the realists' failure to explain what facts are is part of the general failure of modern philosophy to provide a plausible account of the reference of such phrases as 'the existence of flying mammals', 'the existence of prime numbers greater than a million', 'the nonexistence of life on Mars', 'the nonexistence of a greatest prime number', and similar phrases that refer in a specified way to the existence or nonexistence of something. Facts cannot but remain mysterious until it is clearly understood that a phrase like 'the fact that there are flying mammals', if it refers to anything at all, refers to the very same thing as 'the existence of flying mammals'.

The fact that there are flying mammals—namely, *the existence of flying mammals*—makes 'There are flying mammals' true. Similarly, 'that there are no flying turtles' refers to the fact that makes 'There are no flying turtles' true, namely, *the nonexistence of flying turtles*.[18]

The problem of explicating the nature of the facts to which true statements correspond thus transposes to the somewhat more tractable problem of explicating the referents of phrases of the form 'the existence (nonexistence) of *P*'. My approach to this problem rejects Quine's rejection of facts while gratefully adopting two basic Quinian doctrines on existence.

(1) I favor Quine over Putnam on the sameness of meaning of 'exists' and 'true'. Quine not only holds that different expressions differ in meaning unless they are proven to have the same meaning; he also rightly holds that the same (recurring) expression is univocal unless it is proven equivocal (*op. cit.*, p. 131). If so, the burden of proof is on Putnam when he says that 'exists' and 'true' have different meanings in such different sentences as 'It is true that prime numbers larger than a million exist' and 'It is true that brick houses on Elm Street exist'.

---

[18] The identity of the fact that there is (is no) *P*' and the (non)existence of *P*' is clearly not obvious. Quine has no use for facts; he seems not to realize that the fact that there are mammals that fly is nothing other than the existence of flying mammals and that this fact (however referred to) is responsible for the truth of 'some mammals fly'. It is ironic that Quine, who did so much to highlight the centrality of existence in semantics, could so disdainfully dismiss facts. See my article, "Naturalism and Realism," in *Midwest Studies in Philosophy, Volume XIX* (Notre Dame: University Press, 1994), pp. 22-38.

(2) The second crucial doctrine concerns logical form. Going beyond Frege, Quine holds that all elementary sentences, whether they be singular sentences like 'Socrates is wise' or general sentences like 'Every boy admires some astronaut', have the form of existential affirmations or existential denials, and so are parsable as sentences of the form 'Something is (a) K' or 'Nothing is (a) K'. K may be complex as in 'Someone is both Socrates and wise' (canonical for 'Socrates is wise') or 'Nothing is a boy and not an admirer of some astronaut' (canonical for 'Every boy admires some astronaut'). A statement (asserted sentence) of the form 'Something is (a) K (thing)' or '(a) K (thing) exists' expresses the proposition THAT SOMETHING IS (A) K (THING) or THE EXISTENCE OF (A) K; a statement of the form 'Nothing is (a) K' or 'No K thing exists' expresses the proposition THAT NOTHING IS (A) K or THE NONEXISTENCE OF (A) K.

Immanuel Kant rightly taught that existence and nonexistence are never to be regarded as properties of things that are said to exist or not to exist. Having learned from Kant what existence and nonexistence are not, it was left for later philosophers to say what they are. On this topic, modern analytic philosophers have generally followed Frege's lead by distinguishing concepts and objects. Some concepts have objects falling under them, others are empty. According to Frege, existence, while not a property of any object, is a property of the concept under which that object falls. For example, since the concept FLYING MAMMAL has instances, the expression 'the existence of flying mammals' refers to the nonemptiness of the concept FLYING MAMMAL.[17] Other modern philosophers gave answers that were variants of this one. According to Bertrand Russell, the existence of flying mammals is a property possessed by the propositional function 'x is a flying mammal', the property of being "sometimes true." On such views, 'exists' is a second-order predicate whose subject is a first-order predicate. For example, in the formula '$(Ex)(Px)$' we may think of '(Ex)' as if it were predicated of '(Px)', saying of it that it is satisfied or instantiated.[18] In the Frege interpretation, a predicate like 'is a flying mammal' denotes the concept FLYING MAMMAL. '(Ex)(x is a flying mammal)' may then be understood to say that the concept denoted by the predicate has things falling under it; accordingly, 'the existence of flying mammals' refers to this property of the concept: its having things "under" it.

---

[17] See *The Foundations of Arithmetic*, J.L. Austin, trans. (Oxford: Blackwell, 1952), section 53. For an excellent discussion of the doctrine that existence is an attribute of concepts, see C.J.F. Williams, *What Is Existence?* (New York: Oxford, 1981), chapter II.

[18] Quine gave this view its crispest formulation when he said that "to be is to be the value of a variable."

The old-fashioned correspondent realist does not find these "second-order" conceptions of existence technically wrong so much as altogether missing the point as explications of phrases of the form 'the existence (nonexistence) of *P*'. The classical realist wants a "robust" account of what the referring phrase 'the existence of black holes' refers to. We begin with the truism that the existence of black holes is a feature of the universe. That the property of being nonempty or having instances belongs to the concept BLACK HOLE in virtue of the existence of black holes is quite irrelevant. First, the property belonging to the concept is not that of existence. Second, the suggestion that the existence of something like a black hole is at bottom a property of concepts or propositional functions must be rejected by anyone with the "robust sense of reality" that Russell demanded of an acceptable theory.[19]

III

I now present the version of correspondence realism that I believe offers no purchase to objections of the kind Putnam and Davidson raised in their Dewey Lectures. What I have to say should not be regarded as an attempt to prove that correspondence realism is the correct theory. The primary aim is to disperse the darkness—so congenial to foes of traditional realism—that shrouds the ontological relatum of correspondence.

All truth claims are existential and each is made with respect to some domain under consideration. In what follows, I use 'DC' as an abbreviation of "domain of the claim" or "domain under consideration." The DC of 'There are prime numbers greater than a million' and 'There is no greatest prime number' is the domain of natural numbers. When I assert 'There are flying mammals' or 'Black holes exist', the DC is the external world. The DC of 'Some carnivorous dinosaurs inhabited the Dakota region' is a past world. The DC of an utterance may be relatively restricted, consisting, for example, of the objects in my visual field, as when, looking into a drawer, I say 'There is no hammer here'. The statement 'There is no flying kangaroo but there is a flying horse' may count as true if the DC is "The world of Greek mythology."

---

[19] Frege is immune to the objection that second-order conceptions of existence fail anyway because the existence of things like dinosaurs antedates the formation of concepts. According to Frege, the concept DINOSAUR is "eternal," "antedating" language and people. The same applies to the most original thought anyone ever expressed. The doctrine that concepts and thoughts are eternal is forced on Frege, however. For he cannot both say that the existence of dinosaurs is a property of DINOSAUR and allow that the concept came into being in the nineteenth century.

A domain is a totality existentially characterized by the presence (existence) of certain things and by the absence (nonexistence) of certain things. The positive and negative existential characteristics of the DC are its truth-making facts. For instance, the presence of Francis Crick is a characteristic of the contemporary world and this fact makes 'Francis Crick exists' true. The nonexistence of United States senators who are not citizens makes 'Every United States senator is a citizen' true.

When Wittgenstein, pointing to a chair in his kitchen, says it is blue, the fact that the chair is blue is the presence (in his kitchen) of the blue chair and that existential characteristic of the DC is "the reality" to which 'This chair is blue' corresponds.

Call a given domain "{*K*}ish" (pronounced "kayish") if it has a *K* thing in it and "un{*K*}ish" otherwise. For example, the domain of natural numbers is {even prime}ish and un{greatest prime}ish. {Even prime}ishness—the existence of an even prime number—and un{greatest prime}ishness—the nonexistence of a greatest prime number—are positive and negative existential attributes of the domain that make true such statements as 'There is an even prime number' and 'There is no greatest prime number'. The DC of a great many statements is the (real, contemporaneous) world. Elves do not exist but elks do. The world is {elk}ish but un{elf}ish, {Hilary Putnam}ish but un{Santa Claus}ish. We accept that the existence of elks is no more an attribute of elks than the nonexistence of elves is an attribute of elves. But we nevertheless maintain that elk-existence and elf-nonexistence are attributes; {Elk}ishness—the existence of elks—and un{elf}ishness—the nonexistence of elves—are attributes of the world.

As existential characteristics of the real world, facts are real. Nevertheless, speaking strictly, they do not exist. What exists is present in (exists in) the world, but while the world is real, it does not exist, nor do its properties. My cat Daisy exists as a constituent of the totality of things that constitute the real world but her existence, while it characterizes the world, is not another constituent of the world. (Daisy is present in the world, her presence is not.)

Facts are sometimes called states of affairs. That misleadingly connotes 'situations' which, in turn, connotes something locatable in the world. But facts are not situations and cannot be found in the world. If facts are to be talked of as "states of affairs," we should think of them as states of (not in) the world.[20] Strawson was right to say that facts are not in the world; he was wrong to conclude that this meant they could not possibly be counted as the objective relata of correspondence.[21]

The properties of things in the world are ordinary properties; the properties of the world are existential properties. Just as the terms 'philosopher' and 'American' denote Putnam while signifying different characteristics (being a philosopher, being an American) that Putnam possesses, so the statements 'There are American philosophers' and 'There are no Zulu astronauts' denote the world while signifying different characteristics (the existence of American philosophers and the nonexistence of Zulu astronauts) that the world possesses. More generally, the variety of true statements made with respect to any given DC will signify different facts that are existential attributes of that DC, but all will denote the DC itself. Thus, all true arithmetical statements denote the domain of numbers and all true statements of physics denote the totality of things that constitute the physical world.

This gives us two ways of understanding the truth of a statement:

A statement that is true is a statement that signifies a fact.
A statement that is true is a statement that denotes its DC.

---

[20] What exists is a member of the domain under consideration (DC). We often make statements whose DC consists of facts. For example, if someone says 'Elves do exist' (with the real world, *W*, as the DC) you might counter with 'There is no such fact as the existence of elves'; the DC of your statement is a domain, *D*, consisting of the facts that characterize *W*, your claim being that *D* is not characterized by the presence of any such fact as {elf}ishness. The facts about *W* exist in *D*, not in *W*. Sometimes the DC is hybrid consisting of things in the real world and the facts that characterize the world. Thus, when I (truly) say "'Lions exist' corresponds to the fact of the existence of lions," the DC has facts characterizing the world as well as sentences in the world as constituents. What I say is true because of the presence in this DC of a sentence (namely, 'Lions exist') to which a fact (namely, the existence of lions) corresponds.

[21] Moore, an old realist who thought of facts in terms of existence, was sensitive to the "nonexistence" of facts, noting that "it seems unnatural to say...that the fact that lions exist itself exists. It is quite natural to say of a lion that he exists, but it does not seem natural to say that his existence exists"—*op. cit.*, pp. 311-12. Although facts are not in the world, propositions, including FACTS, are in the world in the way symphonies and other human products are. I came to see that facts are not in the world before reading these parts of Moore. He anticipates me here as well as in the matter of the equivocal use of 'fact' and expressions of the form 'that *p*'.

False statements express false characterizations but they are doubly vacuous: they fail to signify facts and fail to denote the world.[22]

Putnam says that the truth of 'Lizzie Borden killed her parents with an ax' "depends on *how things were back then*" (501). Although he speaks of truths as being dependent on states of affairs, he has no use for Moore's question concerning the nature of states of affairs ("how things are"); for that question accepts the assumption that facts or states of affairs have some common nature as truth makers, the assumption, as Putnam put it, that "there is a 'substantive property' [that] underwrites the very possibility of using the word 'true'" (502).

The metaphysical realist does indeed hold that a "substantive property" underwrites the truth of the claim that Lizzie Borden killed her parents with an ax. The domain of that claim is the nineteenth-century world in which Lizzie lived; and the property in question is an existential property of the domain: its {Lizzie-Borden-killing-her-parents-with-an-ax}ishness. The existence of the ax-wielding parricidal Lizzie makes 'Lizzie Borden killed her parents with an ax' true of that world.[23]

Having presented his own version of what purports to be a correspondence-free "common-sense realism," Putnam asks us to "appreciate the enormous *differences* between common-sense realism and the elaborate metaphysical fantasy that is metaphysical realism—the fantasy of imagining that the form of all knowledge claims is fixed once and for all in advance" (514). The metaphysical realist's rejoinder to this is: "All knowledge claims do have a common form: 'The existence (nonexistence) of *P* is a characteristic of the domain'." Because '*P*' is a place holder for all kinds of terms (from proper names like 'John Major' to terms like 'salinity', 'prime number', 'prime minister', and 'boy who envies every astronaut') and because the domains under consideration in knowledge claims are endlessly various, nothing except the existential *form* of all knowledge claims is fixed in advance.[24]

---

[22] Let [*s*] be the proposition expressed by an elementary statement, '*s*', and let <*s*> be the fact that makes [*s*] true. Then '*s*' signifies <*s*>, which is said to correspond to [*s*].

The relations between statements, the propositions they express, the facts they may signify, and the "world" they may denote provide several alternative and equivalent ways of formulating the truth conditions of '*s*':

'*s*' is true if and only if [*s*] characterizes the world.
'*s*' is true if and only if [*s*] is true of the domain.
'*s*' is true if and only if <*s*> is a characteristic of the domain.
'*s*' is true if and only if [*s*] corresponds to <*s*>.
'*s*' is true if and only if '*s*' denotes the domain.
'*s*' is true if and only if '*s*' signifies a fact.
'*s*' is true if and only if '*s*' signifies an existential state of the world.
'*s*' is true if and only if '*s*' expresses a true proposition (a FACT).
'*s*' is true if and only if [*s*] is a FACT

Compound statements have the familiar truth conditions. Thus '*r* & *s*' is true if and only if '*r*' is true and '*s*' is true, and so on.

[23] To avoid repeated qualification, I follow Putnam in speaking of Lizzie Borden as if she were a proven murderess; in fact, she was never convicted of the crime.

It is worth noting that Putnam's denial that knowledge claims have a common form comports with his view that 'existence' is not univocal in such phrases as 'the existence of brick houses on Elm Street' and 'the existence of prime numbers greater than a million' (450fn.). But as Quine correctly observed, there is no reason whatever to say that 'exists' (or for that matter, 'true') means different things just because we happen to be talking about very different things (*op. cit.*, p. 131).

IV

We have been comparing Putnam's realism to the "metaphysical" realism he never ceases to deride. The old realist would have little quarrel with Putnam were he not so quarrelsome. Of course, 'Caesar crossed the Rubicon' makes us aware of the fact that he did just that. The pivotal question, as Moore saw, is: What is a fact? What kind of feature of reality is it? Recent philosophers (including the last two Dewey lecturers) were too down on facts and correspondence to take this question seriously. Davidson is characteristically dismissive: "[T]here is nothing interesting or instructive to which true sentences might correspond" (*op. cit.*, p. 303).[25]

I have pointed to some of the hidden costs of treating facts as *entia non grata*: the deflationist fragmentation of truth, the "second-order" conceptions of existence. It is more to the point to say that the lack of an adequate conception of existence is not so much a cost as a root cause of the mystification of facts. When philosophers do not get existence right, the true nature of the ontological relata to which truths correspond is veiled.[26]

---

[24] Despite its comprehensiveness, the purely existential form of truth-making facts does not cover ethical truths. But it is possible that ethical truths could be explained by existential features that make the world better or worse. A theory of ethical truth along realist lines might include the kind of mondial attributes God observed in the created totality he surveyed when he "saw it was good."

[25] In this matter, Davidson expresses the view of many contemporary philosophers from Strawson to Putnam. This fashionable *non grata* attitude to truth-making facts is quite unwarranted.

[26] For example, Moore, a correspondence realist, who believed strongly in truth-making facts but who was in thrall to Russell's and Frege's "second-order" way of thinking about 'exists', had to "confess" he did not know what facts were.

We lift the veil when we correctly answer the question inadvertently left us by Kant: "If existence is not an attribute of the things that exist, of what is it an attribute?" The "robust" answer is that specified existence and nonexistence (for example, the existence of lions, the nonexistence of griffins) are mondial attributes, attributes that belong to the world in virtue of what is in it as opposed to what is not. A statement like 'Some tigers are tame' corresponds to (is "made true" by) the existence of tame tigers, a positive attribute of the world; 'No tigers fly' corresponds to the nonexistence of tigers that fly, a negative attribute of the world. Here, too, we have the answer to Moore's question about the nature of facts: the world's positive and negative existential properties are its truth-making characteristics—the facts or states of affairs—that underwrite our knowledge claims.

Davidson has rightly said that "realists' views of truth...must be based on the idea of correspondence." Since he agrees with Putnam (and with Frege, Strawson, and Quine) that correspondence "cannot be intelligible," Davidson eschews the realist label. Putnam, on the other hand, claims that one can abjure correspondence and still be a realist. Of wider consequence is the shared belief that classical correspondence-to-fact realism is hopelessly incoherent. That verdict, which is without warrant, has blighted realist philosophy for the better part of this century.

FRED SOMMERS

Brandeis University

THICK EPISTEMIC ACCESS: DISTINGUISHING THE MATHEMATICAL FROM THE EMPIRICAL*

One of the little tragedies of my early childhood is that I never saw a hobbit; and because I did not, I regretfully had to conclude that there were no such things. Along with Santa Claus, mermaids, and Narnia, hobbits eventually had no place among my ontological commitments—and this despite what others (kindly adults, deluded playmates) persisted in telling me.

Apparently, I was not alone. There is a long tradition of taking observation as the final court of appeal when it comes to questions of ontology. Doing so, of course, makes quick work of the status of mathematical and theoretical terms: they are terminological instruments, necessary to do the work empirical theories do, but otherwise picking out nothing real. Perhaps the desire to enshrine observation as criterial of ontology this way explains the appeal of mathematical nominalism; certainly, that layperson outraged by the philosopher's attempts to convince her of the existence of mathematical objects will often be provoked to point out that the philosopher has apparently forgotten that, after all, one does not touch, see, smell, or *otherwise deal with* mathematical objects.

The same was true of atoms once (before 1900, say). Perhaps, therefore, it is no surprise that, at roughly about that time, Ernst Mach was skeptical of the existence of such things, nor that, in 1877, Pierre-Eugene Berthelot asked (rhetorically): "Who has ever seen a gas molecule or an atom?"[1] So too, I suppose, it is no surprise either that the *vera causa* principle—that one should invoke causes for something only if such causes can be independently *observed*—played such an important part in the history of geology.[2]

---

* My thanks to Arnold Koslow for comments on an earlier version of this. Thanks are also due to the audience at the 1996 American Association for Symbolic Logic meetings where I read that same version. A word about my use of 'mathematical' and 'empirical': I intend to explicate these notions via my forthcoming distinction between items to which we have either thin or thick epistemic access. But some terms, 'center of mass', for example, I take to refer to items to which we have thin epistemic access, even though they might not be described by some as 'mathematical'. I do not discuss such terms in this paper, and do not have them in mind whenever I use the word 'empirical' here.

[1] This is quoted in Michael Nye, *Molecular Reality* (New York: American Elsevier, 1972), p. 7. But I owe the quote to Penelope Maddy, "Taking Naturalism Seriously," in D. Prawitz, B. Skyrms, and D. Westerståhl, eds., *Logic, Methodology and Philosophy of Science*, IX (New York: Elsevier, 1994), pp. 383-407, here p. 389.

[2] For details, see Rachel Laudan, *From Mineralogy to Geology: The Foundations of a Science, 1650-1830* (Chicago: University Press, 1987).

0022-362X/97/9409/472–84

Rather than dwell on the numerous examples available (of giving in to this sort of temptation), I would like to ask: Why is it so tempting to take observation as criterial of ontology? Look to epistemology (one traditional answer counsels us), for the temptation in question is linked not to criteria for *what* exists, but to how we make ourselves *sure about* what exists. Observation is less likely to breed epistemic mishaps than any other method we have for establishing the existence of something.

This is (perhaps) perfectly obvious when it comes to bird entrails, horoscopes, or fortune cookies; but the point is seen as nearly trivial even when the contrast is made with otherwise respectable means of establishing the existence of something. Reliance on testimony or authority, for example, can lead to mishaps, over and above those observation is heir to, because of additional failures in "transmission." For sure, *B* witnessed the murder, but by the time the story got to *C*, the evidence was corrupted by some false inferences on *B*'s part about what she saw, in addition to a few rather significant misunderstandings on *C*'s part about what *B* told him. Similarly, reliance on theory can lead to mishaps, over and above those arising from the observations on the basis of which the theory was confirmed, because theories involve inferences (of one sort or another) from the observational data on which they are based, and such inferences are heir to additional problems that pure observations are not. Although this sort of story is happiest when wedded to attempts to couch observations in theory-independent vehicles (for example, observation sentences), even holists like W. V. Quine clearly find it appealing (recall that they like to describe our beliefs as being located at varying distances from the experiential periphery of our conceptual scheme).

I have recently come to find this traditional counsel misguided, however, and not merely because it involves a (*verboten*) slide from an epistemic premise to an ontological conclusion. My worry is more humdrum: it is that, as many philosophers have noticed, theory often corrects observation. *B* says, on the basis of what she sees through her telescope, that the planets are moving this way and that; *C* says, on the basis of his theory, that they are moving that way and this; and sure enough, when *B* rechecks her observations (cleans a lens, recalibrates an instrument, checks a calculation), *C* turns out to be right. Or (this happened to me recently), I report some observation or other, and because of what the seventeen people around me say really happened, I realize that I could not have seen what I thought I saw.

Now, what is odd (to me) about these kinds of cases is that according to the traditional counsel, observations are more reliable than other sorts of epistemic acts (and historically, this makes understandable the grail-like search for the incorrigible observation, the totally reliable *pure* observation). But the lesson I draw from examples where observations get corrected by other epistemic methods is not the holist conclusion that all observations are theory laden or anything like that;[3] rather, the lesson I draw is that we have missed out on what is special about observation by focusing so single mindedly on its being more reliable (in some sense) than other epistemic methods. That fact, if it is a fact, should follow from other epistemic properties observation has; and in particular, it should be clear on the basis of these other properties of observation why and how it is not always more reliable than other methods, and why it is not when it is not.

I

So let me try again: What is special about observation? By way of an answer, I single out four (significant) aspects of observation:

(1) Observation (epistemically speaking) operates, more or less, independently of what we believe. By this I mean that what we see, for example, is quite robust under a broad range of circumstances. (We blink, and what we see is still there. We walk away, walk back, and it is still there.)

There are several important points to make about this robustness. Notice, to begin with, that our recognition of the robustness of what we observe arises from our ordinary evidential practices, and not from, say, neurophysiology (no claims about the modularity of our sensory faculties are being brought to bear here). If I want to prove the existence of something to someone skeptical of it (moldy bread in the refrigerator, say), and it is possible to point at it, I can do so without further ado.[4]

Next, notice that this robustness of what comes through our senses is compatible with the following facts: we can be wrong about what we think we see; we can be trained to see things we could not see before; we may need a bit of theory in order to understand what it is we are supposed to be seeing, or how seriously we should take what we are seeing. Finally, in qualifying the robustness of what we see with 'under a broad range of circumstances', I mean only to point out that there *are* circumstances—great emotion, drugs, mental illness—where what we see is not robust. The suggestion is not that what comes our way through our sensory faculties is invariably robust; it is simply to point out that quite a lot of what comes through is.

---

[3] I should say that I think *observation*, in a rather strong sense, is theory free, although the language we must couch observations in is not. I give some indications for why I believe this in what follows. But the full story must wait for another time and place.

[4] See G.E. Moore, "Proof of an External World," in his *Philosophical Papers* (New York: Collier, 1962), pp. 126-48.

(2) We have means of adjusting and refining our observational means of access to the thing being seen. I can squint to see it more clearly; I can move in for a closer look; I can reach out and touch it. In these ways, I have resources to help distinguish between what I am "really" seeing and what is only an artifact of my senses (or of how they connect me to the world). How do I tell that "floaters" are in my eye and not several inches in front of me? By various experiments where I determine that the motion of such items is being caused by the movement of my eye (not to mention the fact that a particular floater can be seen with only one eye).

Notice that part of what is involved in our capacity to adjust and refine our observational access to something is the very robustness of that access just spoken of. Indeed, the point of squinting, say, is precisely to make the image (of what it is I am trying to see) more robust.

We can see how these two aspects of observation go a long way toward explaining the datum raised at the beginning of the paper. Robustness is a matter of degree, and one needs to improve on it before one can comfortably rely on what one sees. But robustness is not a matter of "clarity" in any traditional Cartesian sense; for one thing, the point of improving it is not linked to the subsequent veridicality of what we see, but simply to whether or not it is something that can be seen again under similar circumstances. Optical illusions, for example, are quite robust. By contrast, what I see when I am drunk need not be robust, although in the first case the lines are the same length despite what they look like, and in the second, the shadow lurking in the alley is a mugger despite the fact that I only intermittently see it.

(3) Our sensory access to something often enables us to "track" properties of it. For example, I can watch something for several hours (a certain sort of insect, say), and in this way determine a variety of its properties (for example, that it moves at a certain speed, that it can change colors, that it likes moss). Furthermore, I can determine what it does over time: I can construct, that is, an "episodic history" of its actions.

There is a variant of our capacity to track the properties of an object we should also take seriously. Some large objects (such as mountains) are not tracked because, among other reasons, they do not move (relative to us, anyway). But we can still observe their properties by moving around *them*. It is a sort of pun, perhaps, also to describe this as tracking an object; but it is otherwise harmless, I think, to assimilate both these ways of determining the properties of an object over time to the same category.

(4) We can connect certain properties of the objects seen with our capacity to *know* about their properties; for example, given our (partial) understanding of our observational means of access to something, and given what we take certain properties of the seen object to be, we can understand how we are able to recognize, by sight, its color, speed, and so on. This need not be a subtle scientific matter: I can explain easily enough why I can see how fast a flock of antelopes is moving: they are large opaque objects that do not travel very fast (even when panicked). So, too, I can explain my trouble seeing how big certain jellyfish are under certain circumstances: they are (more or less) transparent; but, consequently, I know I shall not have a problem seeing how big they are if I dye them first.

Clearly, just about any respectable scientific theory has posits[5] which are not observational: terms referring to mathematical objects, for example. Indeed, it is practically a kind of category mistake to suggest that mathematical objects are observable (not that certain philosophers have not suggested it anyway). But there are lots of other sorts of posits which are not observable, and yet which are not mathematical either. I have in mind, for example, viruses, electrons, quarks, monarchies, or the Id.

These things, in my view, cannot be observed (despite the extension of the term 'observation' by scientists to certain epistemic interactions with some of these things, and the taking seriously of this extension of usage by certain philosophers).[6] But something almost as good as observation is available.

---

[5] For the moment, the reader can indulge in the Quinean fantasy that scientific theories are couched in first-order terms, and that the existential commitments of these theories are whatever the first-order existential quantifiers are taken to range over. But I should point out that I shall (shortly) consider denying Quinean doctrines about existential commitment. Be forewarned: while Quinean doctrines are being denied, I shall not use the term 'existential commitments' to describe implications of a theory which are in the form of an existential quantifier followed by a formula with one free variable; rather, I shall adopt more neutral terminology: 'quantifier commitment'.

[6] Dudley Shapere, "The Concept of Observation in Science and Philosophy," *Philosophy of Science*, XLIX (1982): 231-67; and Ian Hacking, *Representing and Intervening* (New York: Cambridge, 1983), here pp. 181-83; after him, notice that physicists often say things like, "Of these fermions, only the *t* quark is yet unseen," or "There is no way known other than by neutrinos to see into a stellar interior." In light of what I say shortly, I think it is easy to explain why talk of observation has been adopted here: the epistemic means of access in these sorts of cases also has the four aspects I have attributed to observation. But this, I claim, is no reason to call what is going on *observation*.

Notice that there is nothing about our four aspects of observation which particularly requires them to be aspects of observation. Indeed, the point can be made perfectly explicit with a definition: any form of epistemic access which is robust, can be refined, enables us to track the object (in either sense), and which (certain) properties of the object itself play a role in how we come to know (possibly other) properties of the object is a *thick* form of epistemic access.[7] By this definition, all observations of something are thick.

But how might an epistemic interaction with something be thick without being an observation of it? Here is a recipe. Choose your favorite unobservable object (electrons, say). Then forge *instrumental* relations between yourself and instances of those objects so that our four aspects hold of these relations. That is, design instruments so that: (1) the results of operating these instruments are independent (epistemically speaking) of what the operator of them believes; (2) we have means of adjusting and refining our instrumental access to the thing being seen; (3) our instrumental access to something enables us to track properties of it (in either sense of 'track'); and (4) (certain) properties of the object itself (or instances of the kind of object being studied) can be used to explain how it is we know (possibly other) properties of the object (kind of object) in question.

II

I now want to make a few points about how the four aspects of thick yet nonobservational instrumental access (to something) contrast with the four aspects of observation.

(1) The robustness of any instrumental process is only achieved by linking it to the robustness of observation. If a particular instrumental intervention with something regularly leads to a certain result, this means that a certain class of ordinary observations are invariably gotten when the device is operated. A common tool for recognizing certain acids (in certain forms) is litmus paper, and dipping a bit of it into a liquid can result in a change of color that we can see. So, too, more sophisticated devices may yield flashing lights, certain sorts of visual images, sounds, and so on. Indeed, there is no requirement that any particular sort of intervention must yield a particular sort of observation (a geiger counter makes clicks, something else which is measuring the same thing may produce a photograph); I am only saying that when an instrumental interaction is robust, it achieves this

---

[7] This terminology is derived from my *Metaphysical Myths, Mathematical Practice* (New York: Cambridge, 1994), where I wrote of thick, thin, and ultrathin *posits*. I have come to feel that it is the *epistemic access to* the posit that should be described this way rather than the posit itself, and have adjusted my terminology accordingly.

via a link to observation. This is what enables the robustness of the intervention with a theoretical object to be independent of the theory we have about that object (or even the theory we have about the device being used to interact with the object in question).

(2) Notice that one point of refining our instrumental access to theoretical objects is often robustness and, therefore, the process can go on independently of our theories about the objects accessed or of the instruments utilized (we fiddle to reduce "noise" of various sorts). But crucial to this process of refinement is also the recognition of artifacts of the instrumental process itself (just as with observation), and here theory often, although not always, plays an important role. What we see under the microscope might be an artifact of how we dyed the sample, and what effect that has on light. We may recognize this either because we know (on the basis of a theory) that the item in question does not have the structures we seem to be seeing, or because we can isolate this artifact by using other instrumental means of access to the object under scrutiny. Pure observation is perfectly analogous on this point.

(3) Historically, the capacity to track a theoretical entity instrumentally seems to be an important consideration for whether scientists take its existence seriously or not. Back in 1913, Jean Baptiste Perrin[8] included as part of his important evidence for the existence of atoms, C.T.R. Wilson's experiments which seemed to show the actual trajectory of atoms through an atmosphere saturated with water vapor. So, too, we find the following evidential considerations that "made the 'particle' identification of quarks more secure": "Measurements of the scattering of very high-energy electrons, accelerated by the Stanford Linear Accelerator (SLAC), from neutrons and protons demonstrated that the nucleons contained, or were made of, pointlike charged particles which scattered the electrons strongly. Through analyses of the magnitude of the scattering, which depends upon the size of the charges held by the partons, the partons were identified as the fractionally charged quarks."[9] This is tracking of our second sort.[10]

---

[8] *Atoms* (Woodbridge, CT: Oxbow, 1990).

[9] R.K. Adair, "Quarks," in Rita G. Lerner and George L. Trigg, eds., *Encyclopedia of Physics* (New York: VCH, 1991, 2nd ed.), pp. 1001-04, here p. 1002.

[10] Hacking seems to think that we should become realists about a theoretical object when we start using it to interact with other objects (spraying electrons on niobium balls) (*op. cit.*). But it seems to me that physicists can become committed to something long before that stage and, indeed, that it may be reasonable to become committed to a kind of thing (quarks, for example) even if, due to circumstances beyond our control, we can never use them for anything.

To return to the philosophical themes with which I began this paper, recall that when Quine[11] describes the virtues that a theory (with accompanying posits) should have, he mentions simplicity, familiarity, scope, fecundity, and success under testing. The idea, I think, is that, if a theory has these virtues, we have good (epistemic) reasons for adopting it, and all the posits that come with it. I shall call this sort of epistemic access to a posit, that it is a quantifier commitment (see footnote 5) of a theory which has Quine's five virtues, *thin* epistemic access.

Notice that all posits of a theory, mathematical and otherwise, have Quine's five virtues, when available, divided democratically among them: the only way to deprive a posit of its share of virtue is to reformulate the theory *sans* quantifier commitment to that posit, while still preserving all five virtues. No wonder that on such a view, epistemic reasons for committing oneself to the existence of mathematical objects seem exactly on a par with epistemic reasons for committing oneself to the existence of anything else.

If one were more or less convinced, however, that thin epistemic access is not enough, that ontological commitment requires thick epistemic access, and if one were more or less convinced by the scientific record that scientists really do become convinced of the existence of something empirical when they manage an instrumental interaction with it which has the four aspects that I described observation to have, one might be tempted to treat thick epistemic access as criterial for existence. If thick epistemic access *is* criterial for existential commitment, then (this is a neat payoff!) mathematical objects are crisply divided from empirical ones: they simply do not exist on this view.

Well, not so fast. Before instrumental access to atoms was forged, the atomic theory was quite useful in chemistry. Penelope Maddy describes it as enjoying Quine's five virtues in abundance, and this looks right to me (*op. cit.*, p. 388). But it is not that scientists generally denied or generally accepted the existence of atoms; the matter was controversial. Indeed, it seems perfectly obvious that the matter was controversial because everyone recognized that atoms were too small to observe;[12] and many scientists found the other sorts of (inferential) arguments available to support the atomic thesis too risky.

---

[11] "Posits and Reality," in his *The Ways of Paradox and Other Essays* (Cambridge: Harvard, 1976, rev. and enl. ed.), pp. 246-54, here, p. 247.

[12] I suspect that part of the reason for the centrality that questions of observation played in the early debates about the atomic hypothesis was due to the fact that no one quite realized that there was a possibility of instrumental access (to atoms), possessing more or less the epistemic virtues of observation, or how powerful (epistemically speaking) the forging of such instrumental access would prove to be.

Contra our criterial suggestion regarding thick epistemic access, however, this seems to allow the possibility of a *defeasibility* condition: a theoretical item can be one we think exists, even if we do not have appropriate instrumental access to it, provided that we can tell a decent story, in terms of its properties, about why we cannot have such instrumental access to it. Perhaps it is even an epistemic requirement that one give rough conditions under which the items in question could be ones we have instrumental access to (but I am not sure about this last claim). In any case, without thick epistemic access to something (or to instances of a kind of somethings), we have to rely on other less secure methods of establishing its (their) existence, and perhaps therefore it is no surprise to find controversy in such cases.

Therefore, one may think, the crisp distinction between mathematical and empirical posits does not hold, for a defeasiblity condition applies to mathematical posits, too: after all (one could claim), such objects are not *in* space and time, and that is why no instrumental or observational access to them is possible.

I do not think this last consideration should be taken seriously. It is not as if mathematical objects really are not in space and time, or that that is a genuine *property* of them mathematicians study (along with their other properties). This is just the wrong picture: what is really going on is that one simply does not bother giving a defeasibility story about mathematical posits; one already understands that, of course, they are not the sorts of things one would even want to get in instrumental or observational contact with.

III

I like this last suggestion and shall spend the rest of the paper explicating it. The difference between an empirical posit to which one has thick epistemic access and an empirical posit to which one only has thin epistemic access is that one can (and should) tell an epistemic story about why one can get in observational or instrumental contact with the first posit but not with the other. But being an *epistemic* story, it should not follow that existential commitment is criterial on a difference like this (which is not to say that one might not feel that one knows enough to be sure that an empirical posit, to which there is only thin epistemic access, exists).

Since the distinction between empirical objects and mathematical ones is not epistemic at all—mathematical objects are not the sorts of things to which one should even try to forge thick epistemic access—one can wonder, what kind of distinction is being made here between empirical and mathematical posits? A metaphysical one? If so, what bearing (and why) should it have on the question of whether we are existentially committed to mathematical objects by virtue of needing them for empirical theories? In other words, how does this distinction bear on what has come to be called the Quine-Putnam indispensability thesis?

I have actually thought of an argument *against* the Quine-Putnam indispensability thesis, an argument for why we should not take ourselves to be committed to the existence of mathematical objects just because our (best) scientific theory requires quantification over them. It is this: despite the fact that every posit is treated in the same way, logically speaking, by quantifiers in a theory, nevertheless, mathematical posits get into scientific theories *the wrong way*. Let me explain.

Recall that, if there is an upper estimate on what the numbers of cats and dogs are, you can have a theory that there are as many dogs as cats without committing yourself to the claim that there is some number that is the same number of dogs as it is of cats. By contrast, if an upper limit on the numbers of these things cannot be sensibly given, your theory must commit you to this number.[13]

Here is another illustration from geometry: we live in a universe containing differently shaped objects of varying distances from each other. How do we provide a general scheme for describing the sizes of these objects, and where they are located? If we are only interested in objects in our immediate (finite) locale, we can do this: divide the space in question into a finite number of small boxes of the same size, associate each box with a predicate, and use the predicates to locate the objects by listing the boxes with which they overlap; this also supplies a crude metric that enables us not only to give the volumes of the objects, but also to describe how far apart they are from each other.

---

[13] This illustration, of how one can find oneself committed to numbers, has come up before in the literature; for example, Quine, *Methods of Logic* (New York: Holt, Rinehart, and Winston, 1972, 3rd ed.), p. 241; and Dale Gottlieb, *Ontological Economy: Substitutional Quantification and Mathematics* (New York: Oxford, 1980), p. 42.

Some philosophers try to avoid said ontological commitments by augmenting their logical resources by, for example, introducing infinitely long sentences, or using some set theory, or quantifying over functions, or importing second-order logic. It is not clear that doing so enables one to escape the ontological commitments, but, regardless, going this route already shows that one thinks one has gained commitments to numbers in a special way since one never uses methods like this to avoid ontological commitments to, say, dogs. See the considerations immediately following the second illustration.

I should also add that I am implicitly assuming here that the theory that the number of cats is the same as the number of dogs is not one that postulates a natural 1-1 functional relation between cats and dogs (as there is between people and their noses, say); for in this case, such a function could do the job needed without additional quantifier commitments.

Notice that in adopting this clumsy expedient, and despite my ontologically tendentious way of describing what we did ("divide the space in question into a finite number of small boxes"), we have not incurred any new quantifier commitments. We have gained lots of new predicates of course, phrases of the form '*x* is located at *P*' (where '*x*' is a variable and '*P*' is not).

Such lists of predicates (especially if the boxes are small and numerous) are not very perspicuous nor easy to manipulate; in particular, they do not enable us to recognize (at a glance) similarly shaped objects by their descriptions. Apart from that, any increase in accuracy will require smaller grids and, consequently, an increase of the number of such predicates. Replace the boxes with points in three-dimensional space, label the points with real-number coordinates, and suddenly one has a powerful and perspicuous language for characterizing (*certain*) objects and their locations. By judiciously choosing coordinates, one can describe a particular spherical object, for example, as that object such that for all x, y, and z (where x, y, and z range over points) $x^2 + y^2 + z^2 \leq 17$.[14]

What these two examples show is that in becoming quantifier committed to mathematical entities in these ways, we are not implicitly using an inference to the best explanation, hypothetical-deductive methods, or any other traditional scientific method for establishing the existence of something. We are simply being "driven" to quantify over (particular sorts of) mathematical objects, because the language we use to describe things has to be formulated a certain way if we want greater facility to manipulate descriptions of objects and recognize similarities among them; and indeed, this fact is crisply recognized by standard scientific practice: posits brought into play in this manner are ones to which we never attempt to have thick epistemic access.

---

[14] Notice that what Quine calls "ideology" moves centerstage from here on: since we cannot describe objects by explicitly listing the points at which they are located, we must develop other linguistic tools for characterizing classes of points, and these tools are open ended in character; for example, algebraic formulas enable us to characterize a certain class of figures, transcendental functions enable us to characterize a still larger class, infinite series of functions expand it further, and so on. There are interesting issues to discuss here, but they must wait for another time and place; what is important for my purposes is to notice that the mathematical development described here is not "ontological" in character but "ideological" (once we have admitted quantifier commitment to points and classes of such).

Notice also that, in any case, we cannot describe the geometry of actual objects perfectly, but must approximate them by shapes that we can describe (given the linguistic tools available at a time for characterizing collections of points). This, however, is true of any systematic method for locating objects and describing their geometric properties, including the innocent grid system of predicates with which we started.

Consider my second illustration again: despite the adoption of a geometrical language that takes space to be composed of infinitely many points in a richly structured way—ordered, dense, complete—there is no attempt to forge thick epistemic access to such points, or determine in an instrumental way their purported properties—determine, for example, if they are really dimensionless. Nor does anyone ever bother to give a defeasibility story for why we do not do these things.

Notice what is being suggested here. In each case, we can tell what factors are behind our theoretical need for a quantifier commitment: we can tell whether the factors in question have to do with aspects of the subject matter about which we are formulating a theory (broadly speaking, what we might call facts about the world) or only with facts about how we have to formulate our theories.[18]

Thus, the distinction between posits that require (defeasible) thick access and ones that do not is really a distinction between how posits come into play in a scientific theory or, less grandly, in the language used to describe the world around us. Obviously, the distinction in question is not an epistemic one (that is why one is not required to provide defeasibility stories about why we do not have thick epistemic access to mathematical objects) nor is it a *metaphysical* one (for no attempt is made to show that mathematical objects are or are not different from empirical objects in some profound metaphysical way; from the scientific point of view, the issue is a "do not care").

Finally (and this is really important to stress), although the distinction I am developing here is, broadly speaking, one that marks a border between the effects of "language" and the effects of the world on quantifier commitment, it is *not* a distinction that depends on anything like an analytic/synthetic distinction (at least as that distinction is ordinarily understood). For my distinction is supported by the presence of (defeasible) thick epistemic access to certain posits, and by our recognition of how quantifier commitments to other (mathematical) posits arise from the facility they yield in our capacity to describe things.

---

[18] This last move is, of course, based on just two illustrations; but the illustrations are quite paradigmatic: one shows how counting practices can motivate quantification over numbers, and the other shows how our characterizations of the geometric properties of objects (location, distance from each other) can motivate quantification over points. Nevertheless, I cannot be said to have shown that every case of quantifier commitment to mathematical entities arises through what we might call language-structural considerations (to show that, one would need to look at every case of quantification over mathematical entities). But how the other cases are supposed to go should be quite clear from the examples given.

So this distinction does not rely on considerations of any sort about meaning, or on epistemic considerations about unrevisability (which are two classical sources of support for the analytic/synthetic distinction). Indeed, the distinction I have drawn is, for example, compatible with "meaning holism," or with the various epistemic views Quine and others have propounded over the years about mathematics, empirical science, and their relations. For all that has been said in this paper, empirical considerations may force one to revise the mathematical claims one takes to be true (and this even if it is for reasons of linguistic facility that one finds oneself quantifier committed to certain posits—think of geometry).[16]

Rather, to repeat, I am pressing the suggestion that we can easily distinguish the sources of quantifier commitment as being due to facts either about the world or about how (scientific) languages must be formulated. This distinction, a perfectly commonsensical one, I think, is well worth taking seriously: it is what is behind the claim I sometimes hear physicists make when they describe mathematics as "the language of physics."

A case is not being made here that mathematical objects do not exist (I do not know how to make a case like that); rather, it is only being argued that a need to quantify over mathematical objects in the context of a scientific theory that we take to be true is no reason, by itself, to think that such items exist.[17]

JODY AZZOUNI

Tufts University

---

[16] The point just made is one about which distinctions are independent of which; it is not an implicit acceptance of the claim that, in fact, the various forms of the analytic/synthetic distinction are untenable, or that empirical construals of mathematical epistemology are unproblematical. For my views on these complicated matters, see my *Metaphysical Myths, Mathematical Practice*.

[17] I am writing as if I have given an argument for this position, but perhaps I am only shifting burdens. What response is there to someone who argues, "You have to say, 'there is the same number of cats as dogs', and that is enough of a reason by itself to think that item (the number) exists. Who cares about whether this posit got into the theory 'the wrong way'; the point is that we cannot get it *out* again." What can anyone say to this? (Recall what Ludwig Wittgenstein said about spades and how they turn when they hit bedrock.) I think there is something important going on here regarding alternative criteria for what we take to exist, but this is a matter I must explore elsewhere.

BOOK REVIEWS

The Morals of Modernity. CHARLES LARMORE. New York: Cambridge University Press, 1996. x + 226 p. Cloth $49.95, paper $17.95. *The Romantic Legacy*. CHARLES LARMORE. New York: Columbia University Press, 1996. xvi + 100 p. Cloth $20.00.

These two books are best read as the second and third members of an impressive trilogy, whose first volume was *Patterns of Moral Complexity*.[1] Each has its own unity, though all three are collections—the first two of essays, mostly published earlier; *The Romantic Legacy* (RL) of three lectures delivered at Columbia University in 1994. Each essay or lecture has its own points to make, but each also contributes to the statement and defense of an overall position. There are changes of emphasis in the later restatements and expansions of Larmore's views, but readers of these latest books will often find it helpful to refer to their predecessor.

Larmore's prose is always clear and often elegant. His work has historical as well as philosophical dimensions, and he defines his own positions with reference to French and German debates as well as to those of the English-speaking world. (Larmore himself writes in both French and German and two of the essays in *The Morals of Modernity* (MM) first appeared in his *Modernité et Morale*.[2]) Together, his books constitute a major contribution to moral and political theory.

The nature of modernity, as Larmore understands it, is not to be captured by any simple formula. We have just too many inheritances, some from the Enlightenment, some from Romanticism, and some from still earlier periods, resulting in a peculiarly modern awareness of a variety of heterogeneous and competing goods and a peculiarly modern range of disagreements about those goods. This by now deep-rooted inability to agree would by itself preclude modern acceptance of anything like Aristotle's account of the virtues and of the human good, presupposing, as it does, possibilities of rational agreement. An additional ground for this rejection is the modern recognition that there are "multiple sources of value" (MM 156), both of duties and of goods, and that this multiplicity is part of what engenders unresolvable disagreements. Pluralism about values is therefore one central and distinctive aspect of modernity, and the accompanying multiplicity of disagreements among reasonable people, even about pluralism itself, brings about a realization that "the more we converse, the more we disagree, even with ourselves" (MM 12).

---

[1] New York: Cambridge, 1987.

[2] Paris: Presses Universitaires de France, 1993.

It is not just that Greek virtue ethics is incompatible with modern pluralism and that recent scholarly attempts to represent Aristotle as a pluralist are misrepresentations (on this Larmore is plainly right). It is also that ancient Greek conceptions of reason in general cannot accommodate the facts of reasonable disagreement, so that all projects of returning to some ancient standpoint, whether NeoAristotelian or Straussian or whatever, are bound to fail and do fail. To this it might be said: If Aristotle and the standpoint of modernity are incompatible, might not Aristotle be in the right? Larmore's reply is that philosophy is a historically situated activity. Because "our most fundamental concepts and forms of experience follow the vicissitudes of experience" (MM 9), it is only by attending to the inherited patterns of belief and experience that constitute our own present moral culture that we will be able to "see clearly how we are to go on" (MM 10) from wherever we are. Hence Larmore's claim that "modern forms of experience should *constitute* or be reflected in the very categories of our moral thinking" (MM 2). So no ancient philosopher could answer the questions that Larmore takes to be *ours*.

His conception of philosophy puts him at odds with Enlightenment as well as with Aristotelian conceptions of reason. The Enlightenment belief that there are norms of rational action such that all rational agents must conform to them, and such that their categorical claims upon the allegiance of rational agents amount to what we can recognize as morality, are false. Rationally indisputable norms are too abstract to supply what is needed. This is part of what we need to learn from the Romantics. More than this, Romanticism introduced us to a "sort of self-understanding that *ought* to be ours, namely, one in which we view our deepest commitments not as the objects of autonomous choice but rather as the expression of our belonging to a given form of life" (RL 58). Reason should not be understood as "requiring us to stand back critically from our way of life as a whole, as though its ultimate aim were to view the world *sub specie aeternitatis*," but should instead "be seen as directed towards how we are to go on within the terms of this way of life" (RL 58). We begin, that is to say, from our existing beliefs and we do not need to justify those beliefs, until and unless we have been afforded good reason to doubt their truth; and a good reason is one that is supported by our own other beliefs. To justify a proposition is "to give reasons that dispel a doubt" as to its truth by solving the particular problem presented by that particular doubt. "Not belief itself, but change of belief, forms the proper object of justification" (MM 60).

We are to begin then in morality as elsewhere with the beliefs that, as inhabitants of modernity, we cannot but acknowledge as our own. Among these is belief in a morality of universal categorical imperatives, "a set of duties binding on all" (RL 37), according to which the right is prior to the good. The use of a Kantian idiom is apposite. For although Larmore rejects Immanuel Kant's conception of practical reason in the *Groundwork for the Metaphysics of Morals*, he takes himself to be following the Kant of the *Critique of Practical Reason* not only in his characterization of moral principles, but also in his view that the "categorical or unconditional character of moral duties" is such that "we are immediately aware of it in our conscience without being able to justify it by means of anything else" (RL 30). How do we become thus aware? Larmore's answer, that it is by being adequately socialized into the form of life to which these beliefs are central, is of course not Kant's. The distance between his position and Kant's is further emphasized by his insistence on the occurrence of conflicts not only between a variety of duties, ideals, and conceptions of the good, but also between consequentialist considerations and deontological convictions (RL 157-58).

How are such conflicts to be confronted? Not by any appeal to systematic principle (on this see *Patterns of Moral Complexity*, chapter 6). Because there is no single standard of value, rival claims are often incommensurable. But Larmore claims that "values can prove comparable without being commensurable" (RL 160). Here, he follows Herder, both generally in affirming that rival sets of values may be comparable, and more particularly in asserting the superiority of moral universalism to its rivals. Our weighing of heterogeneous goods "can give us reason to believe that some such goods are more important than others" (RL 162).

How does such weighing yield such conclusions? "I admit that I have no fully satisfying answer to this question" (RL 162). Larmore does, however, rule out certain types of answer. We are not, for example, to assign weights to different and rival values by considering the place that each has in our life plan, in some "unified conception...of our overall purposes and the way to achieve them" (RL 95). For one of the Romantic values that contributes to our modern sense of individuality is authenticity and to be authentic is always to be open to the unplanned. "Life is not properly the object of a plan" (RL 95). Yet Larmore is committed to holding that as reflective agents we have the resources needed to justify our moral claims, since on his view there is such a thing as moral knowledge, which is "best understood as the reflective knowledge of reasons" (RL 96).

He is, therefore, also committed to denying both all noncognitivist accounts of moral judgment and all those cognitivist accounts which make the objects of moral knowledge something other than reasons. His principal noncognitivist targets are Gilbert Harman, J. L. Mackie, and David Hume; and the cognitivisms that he rejects include John McDowell's reliance upon an analogy between moral judgment and perceptual judgment and the cognitive naturalism of Richard Boyd and others. Where these fail, cognitivist and noncognitivist alike, is in their inability to give an adequate account of those normative truths about reasons in which moral knowledge consists. Noncognitivism, even in the sophisticated forms given to it by Simon Blackburn and Alan Gibbard, cannot recognize the independence of our preferences and interests that genuine norms possess; and no consistent naturalism can find a place for such norms.

The core morality that is at once the expression of distinctively modern forms of life and the object of our moral knowledge supplies us with reasons for holding two principles that jointly constitute a defensible liberalism. These principles are to govern public debate among those who as a matter of fact share a common life. The first is a norm of rational dialogue. On matters of disagreement rational dialogue proceeds either by retreating to the neutral ground of shared belief or by bypassing the disagreement and solving the problem on the basis of this common ground. The second principle is a norm of equal respect for persons, a principle that requires that in political life we do not secure compliance to the solution to some problem only by force, but are always prepared to go on talking. Force may indeed sometimes be necessary, but "no one should be made by force to comply with a norm of action when it is not possible for him to recognize through reason the validity of that norm" (RL 220). These two norms jointly define the neutrality that liberalism requires of government. It is in the light of these principles that Larmore objects both to Carl Schmitt's hostile critique of democracy and to Jürgen Habermas's ethics of discourse.

It will be clear, even from this bare summary of Larmore's moral and political philosophy, that there is something in it for almost everyone to quarrel with. Aristotelians, Humeans, Kantians, Nietzscheans, Straussians, expressivists, constructivists, and a number of others will all find themselves identified as the authors of errors. To each it will doubtless seem a legitimate complaint that their views—or at least Larmore's interpretations of their views—are often given dismissively short shrift. But such complaints, though sometimes just, may obscure Larmore's achievement, that of stating in a tolerably extended version a set of systematically related positions which it is important to confront. Had he given all his opponents their due, he might have needed more than an academic lifetime. How then should we proceed critically? By framing questions to which Larmore does not as yet provide anything like adequate answers.

A first set concerns Larmore's conception of knowledge. On Larmore's account, we can have knowledge of the existence of reasons. He calls his account of reasons Platonistic, but the Platonism is given no content, except that it entails a denial of naturalism, itself inadequately defined. On any plausible account of knowledge, we can distinguish between that which we merely take ourselves to know and that which we in fact know. How does Larmore draw this line? His thin Platonism suggests that whether or not there *is* in fact a reason for someone to judge or to act in this way rather than that is independent of that individual's supposing herself to have such a reason. If so, how do we test our beliefs about reasons? The word 'reflection' functions as an uninformative place filler here and it needs to be replaced by a detailed and substantive account; and the account of comparability is much too brief and compressed. Knowledge, whatever else it may be, is that which is able to withstand critical testing. How then, on Larmore's view, do we test our beliefs about reasons?

A second not unrelated set of questions arise from Larmore's singular lack of interest in the genesis of reasons and of beliefs about reasons. We know a good deal about how such beliefs are produced and on occasion manipulated; and we also know something about how some people learn how not to raise questions about their own beliefs, how not to find reasons for changing them. But surely what regard we pay to someone's avowed reasons for acting or judging ought to be affected by this knowledge of the effectiveness of ideological manipulation and of self-deception. Does it follow from Larmore's account of reason that such people are justified in continuing to hold their beliefs? It does seem to follow. Yet *we* should have learned by now, perhaps from Karl Marx and György Lukacs, perhaps from Friedrich Nietzsche and Michel Foucault, perhaps from Sigmund Freud and Jacques Lacan that each of us has reason in certain areas to put ourselves to the question, if we are to acquire, through rigorous testing, the kind of self-knowledge that can deliver us from phantasy and illusion. Our existing beliefs, contrary to Larmore's thesis, sometimes betray their need for justification by our not having found reasons to put them in question.

Larmore's lack of interest in the genesis of reasons is a counterpart to the abstract character of his presentation of the individual-with-her-or-his-reasons as the basic unit of political life. Abstracted from social role and institutional setting, from all those contexts in which character is formed or deformed through those social relationships that distribute power unequally, individuals appear as neither victims nor victimizers and equality of respect for persons as understood by Larmore seems to entail that each is always equally entitled to participation, and a hearing in political dialogue. Does it?

Finally, there are questions about Larmore's use of the notions of pluralism and of multiple disagreement. It is part of the distinctiveness of Larmore's position that he characterizes the morality of modernity as possessing a degree and kind of complexity and variety that is incompatible with most accounts of morality. His account of the reasonableness of agents has no place for the view, held, for example, by Henry Sidgwick, that rational agents with appropriate information will tend to converge in their conclusions. For any such position the questions arise: How are we to distinguish defensible complexity and variety from indefensible eclecticism and incoherence? Do we not need some stronger notion of rationality, if we are to be able to distinguish those moral positions which make serious claims upon our allegiance from those that do not?

Suppose that Larmore is unable to supply satisfactory answers to these questions: Would that deprive these books of their philosophical and moral interest? Not at all. For what Larmore has presented to us is an insightful and illuminating version of the moral and political self-image of our society's dominant culture. A hostile critique of that self-image might understand it as disguising a situation whose moral eclecticism and incoherence are such that fundamental moral and political debate cannot provide its politics with a common mind, and that the outcomes of debate are often in fact determined not by how the arguments go—the incommensurability of standpoints sometimes ensures that they could not be so determined—but by the solicitations of power and interest. Whether such a critique is or is not justified depends in key part on how far Larmore or others are able to supply compelling answers to the critical questions that I have addressed to his work. That by itself makes his books more interesting than most.

ALASDAIR MACINTYRE

Duke University

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 10, OCTOBER 1997

LIBERALISM, SANCTITY, AND THE PROHIBITION OF ABORTION*

May a liberal state outlaw abortion in order to protect the lives of early-stage human fetuses? May it do so without relying on the controversial claim that early fetuses have rights that abortions violate? I have a growing worry that opponents of abortion may be in position to construct a constitutional argument to that effect. This possibility has been opened by Ronald Dworkin's proposal of a new moral paradigm in which he hopes the abortion debate in America will henceforth be conducted—a paradigm designed to secure precisely the constitutional right that I believe it jeopardizes.[1]

I. THE INVESTMENT MODEL

Dworkin believes that most people engaged in the public debate about abortion do not really disagree about the metaphysical standing of the fetus or about whether abortion is murder. Rather, what Dworkin finds most striking is what people agree upon: liberals and

---

* I am much indebted to audiences at the University Center for Human Values/Princeton University, the Center for Biomedical Ethics/Stanford Medical School, and University College/University of Oxford. For comments on earlier versions, I thank Ronald Dworkin, Amy Gutmann, Gil Harman, Steve Macedo, Susan Okin, Nancy Rosenblum, Alan Ryan, Amy Seybolt Tomasi, and my students at Stanford University and Brown University. I am especially indebted to Bernard Williams—for many wonderful conversations, and for encouragement, throughout the course of this project.

[1] See "Unenumerated Rights: Whether and How *Roe* Should be Overruled" (henceforth UR), in Geoffrey R. Stone, Richard A. Epstein, and Cass R. Sunstein, eds., *The Bill of Rights in the Modern State* (Chicago: University Press, 1992); his *Life's Dominion: An Argument about Abortion, Euthanasia, and Individual Freedom* (LD) (New York: Knopf, 1993); and "Tyranny at the Two Edges of Life: A Liberal View" (TTE), *New Perspectives Quarterly*, XI, 1 (Winter 1994).

0022-362X/97/9410/491–513

conservatives agree that every human life, in whatever stage or condition, is in some way *intrinsically valuable* or *sacred*.[2]

Dworkin calls the instrumental value a life has, for the person whose life it is, its *personal value*. It is personal value we aim to protect when we recognize a person's right to life, which is why the abortion debate traditionally centers on the question of whether a fetus is a person. But if we treat the life of any human organism, including a fetus, as having intrinsic value—as a product of wondrous and respectworthy creative processes—then abortion remains morally problematic, even if we decide (as Dworkin says we must) that an early-stage fetus does not have an interest in its own life (LD 73).[3]

Dworkin says that while (most) conservatives and (most) liberals agree that abortion is always morally problematic, they think it worse on some occasions than others. This suggests that both groups view the badness of a waste of life as a matter of degree. What measure of badness do they use when making such judgments?

To find out, Dworkin considers questions about the value of life on which conservative and liberal opinions converge. Most people think that the death of a one-month old fetus is a bad thing but, from the impersonal perspective, that the death of a two-month old fetus is somehow incrementally even worse. Similarly, while people find the death of an infant tragic, they find the death of an adolescent even worse. But regarding the badness of death at the other end of life, toward old age, people's opinions seem to be reversed: people typically treat the death of a ninety-year old, *cæteris paribus*, as less tragic than that of a fifty-year old. From the perspective of intrinsic value, Dworkin says most people think that a human life has a certain natural shape: people treat premature death as increasingly bad from conception to late adolescence, as plateauing through middle age, and then sloping down toward extreme old age.

---

[2] Of intrinsically valuable things, Dworkin says some (species of plants or animals) are the product of *natural* processes (natural selection or, on some views, God's work); other things with intrinsic value (paintings or sculptures) are products of *human* creation (art or culture). Some intrinsically valuable things, such as knowledge, are *incrementally* valuable: the more, the better. But other such things—an animal species, a particular painting, a human life—are valued only when and because they have come to exist: Dworkin calls these *sacred*. Some sacred things (national flags) have their value because of what they are *associated* with. But other things (great works of art; each human life) are sacred by virtue of their manner of creation, their genesis. Of these *genetically* sacred things, people feel a special honor for each human life because each human life is the highest expression of both bases of the sacred, natural and human (LD 83).

[3] Dworkin argues that only things with interests can have rights, and only things that have or have had mental activity can have interests, so he thinks an early-stage (up to twenty-two weeks) fetus cannot have rights (UR 403).

Can we explain these views? The crucial move, according to Dworkin, is to think of the creative inputs in people's lives, whether from nature (natural selection or, on some views, God's work) or from humans (culture and training), not simply as contributions but as *investments*. Looked at this way, people's views about the shape of life make sense: people see any premature death as being incrementally more tragic as more investment is made in the life, and less tragic as that investment is fulfilled.

Dworkin thinks the investment metric also can explain the patterns in which conservatives and liberals disagree about the impersonal value of life. Dworkin asks: Is premature death always, inevitably, a more significant frustration of investments in life than other forms of failure? Decisions about abortion often raise that question. Dworkin says that conservatives, because they place greater importance on natural (divine) investments, tend to answer in the affirmative; liberals, who typically place more importance on human investments, in the negative.

So the abortion debate comes to this: "Is the frustration of a biological life, which wastes human life, nevertheless sometimes justified in order to avoid frustrating a human contribution to that life or to other people's lives, which would be a very different kind of waste? If so, when and why" (LD 94)?[4] Unlike the rights-based model, the investment model leaves it open, morally, whether aborting or not aborting best respects the intrinsic value of life: it all depends on the relative weights different people incrementally assign to natural versus human investments. Since the criteria for assigning weight on these matters come from each person's deepest understandings about the meaning and place of human life in the universe, decisions about abortion are essentially religious decisions. Thus, Dworkin believes his model opens the possibility for constructing a new constitutional argument about abortion, an argument defending a woman's right to procreative autonomy which is based in the freedom of religion clauses of the First Amendment (we shall consider that legal argument below).

---

[4] Compared to the rights-based model, Dworkin thinks the investment model can better explain the view of conservatives who think abortion is permissible only to save the life of the mother: since nature's investment will be frustrated by the death of either, the choice in the mother's favor would at least avoid the additional frustration of the personal and social investment in her life. It also makes better sense of the exception that some more moderate conservatives allow for cases of rape and incest: since conception in such circumstances is "unnatural" in a way that violates God's commands, it seems proper to them to frustrate pregnancies begun like that (LD 94-97). Dworkin says the investment model also better explains the liberal position: liberals think abortion is always a weighty moral matter because they feel an awe at the marvelous creative investment represented by every fetus (even atheists can marvel at the fantastically complex process a developing fetus represents), even when they conclude that abortion is permissible because of the special importance they give to human efforts and investments.

II. AN EPISTEMIC BLOCK

No doubt there are a number of criticisms that adherents to the justice-based paradigm might make of the moral model Dworkin proposes. Still, I believe that Dworkin's investment model has remarkable resonance and potential power. The worry I have about Dworkin's proposal arises from *inside* his model. It begins with an epistemic problem.

Investments, as such, are toward something; if they were not, the notion of frustrating them would make no sense. With a human creative investment in an intrinsically valuable thing—like that in a painting—we can know what the investment is toward: it is toward there being a thing of complexity and wonder, an expression of the artist's creative vision, in the world. We know this, it seems, because there is intentionality attendant to human creative contributions. It seems to be intentionality, however murkily, that makes contributions investments. When it comes to evaluating whether some event (a gallery showing; a studio fire) should count as realizing or as frustrating such investments, we necessarily consult (as best we can) with the terms imposed by that intentionality. We ask how well or ill that event meets with those terms. But with a purely natural creative investment—like that in a species of tree—we are more frankly baffled by what such investments are toward.[5] When the terms of the natural contributor's intentionality are beyond us—whether because of the mysteries of God's ways, sociobiological underdetermination, or old-fashioned evolutionary agnosticism—people are less able to state for themselves when and why actions or events should be counted as realizations or as frustrations of investments.

With a mixed investment—like that in a human life—the epistemic situation regarding frustration or fulfillment is problematic in a special way. For with such mixed investments there is a danger that our confidence in claiming to know the shape of the human side of the investment (the side that aims at a successful career, a fulfilling life) may let us slip into claiming to understand the shape of the whole investment. Perhaps this false confidence arises because the timings of human investments (and return-takings) self-consciously take place within the parameters of natural processes of physical and mental growth and decline. Despite even the fact of biological decline, however, there is nothing to warrant the assumption that the natural and the human investments are toward the same thing: the patterns of what each is investing toward may not map onto each other; or very possibly, what nature is "investing toward" may resist patterning altogether.

---

[5] Notice that uncertainty of this sort is compatible with the conviction, common among religious conservatives, that God always invests toward "the good"—however unknowable the shape of each of His investments.

Anyone who thinks of nature's contributions as investments must admit that he is blocked from knowing what those investments are toward, so he cannot know the shape of the whole investment. Even if Dworkin is correct about the view conservatives and liberals share concerning the shape of a normal human life, the interpretation Dworkin gives of the investment model is unable, by itself, to explain why they have that view. Indeed, if the full spectrum of people's beliefs about such matters actually were captured by Dworkin's weight-based account, then people should have a very different view about the shape of a normal human life: deep uncertainty.

But if a strict reliance on Dworkin's account of the investment model would render people deeply uncertain about nature's investments, why are people not uncertain? Indeed, why are they particularly confident regarding what counts as respecting or disrespecting nature's investments in many wholly natural things? When human projects, directly or indirectly, threaten an entire species of animal (consider the ivory hunting of the African elephant, or the eradication of the snail darter by the Tellico Dam) most people are quite sure that this is a terrible thing, an act of desecration. How do people come to feel so certain about this? Why do they not think that human projects that result in the extinction of a species merely represent, say, one possible way of realizing a return? The certainty people feel in cases where Dworkin's account would predict their uncertainty must be traceable to some further, undiscovered principle at work in their thoughts. What is that hidden principle?

On Dworkin's interpretation of his own model, whether describing how conservatives and liberals agree or disagree, people's measurements of cosmic badness are based directly on calculations of the relative weights assigned to natural versus human investments. But is it true that for people concerned about intrinsic value that this direct calculation about weights is all that counts? Let me suggest that for many such people—conservative and liberal alike—it is not. Dworkin's weight-based interpretation of the model fails to capture a central, further feature of people's beliefs: a rule that both liberals and conservatives employ when deciding what actions are permissible with regard to intrinsic values. Especially when considering natural investments, when people care about "the way the story goes" (Dworkin's phrase), they evaluate what happens in the story primarily by asking who in the story makes it happen. The rest of this essay will be about that idea and the puzzles that attend it.

III. THE INVESTMENT CONNECTION

There are many questions about the notion of investment upon which Dworkin bases his model. We considered one when raising the epistemic problem: the question of towardness. But there is a more pressing question if Dworkin's model is to be applied on the practical level: If these contributions are investments, who controls the passbook? That is, still metaphorically for now, who makes deposits and who makes withdrawals? Who opens the account and who can close it? But, for many people, there seems to be a broad normative presumption that there be some appropriate connection between the first and the second party in each question. People tend to think there should be some appropriate connection between whatever party opens (or makes deposits into) an investment account and the party who withdraws from (or wants to close) that account. If I am regularly depositing money into my savings account, but someone I do not even know is allowed to make whatever withdrawals he pleases, then something has gone wrong with this investment system. The idea (rough for now) that there needs to be some appropriate connection between the identity of depositor and of withdrawer is what I shall call the *investment connection* (IC).

My suggestion is that many people (in their beliefs, if not in their explicit thinking) treat the structure of any sacred investment account—like the structure of more familiar financial accounts—as having a certain integrity. That integrity is violated when there is an inappropriate connection between depositing/opening parties and withdrawing/closing ones. IC is a normative principle. It tells people how they ought to act by serving as an epistemic indicator for determining what actions respect or disrespect sacred investments, an indicator so basic that people treat it almost as an independent value. While I believe IC has impact on the abortion question, note that it is a neutral principle: accepting it does not presuppose either a conservative or a liberal outlook on that question. If one accepts Dworkin's fundamental idea that creative contributions be thought of as investments, one has reason to accept the principle of IC as a guide to what actions may appropriately be taken with investment accounts. As we shall see, many liberals and conservatives who accept (or have beliefs captured by) Dworkin's investment model do accept that principle.

Especially when confronted with the epistemic block I described, people treat an action as violating or disrespecting a sacred investment when that action severs IC. But different sorts of investments require different standards of what counts as "an appropriate connection" between depositors and withdrawers. We can distinguish stricter and looser interpretations of IC, depending on whether/which proxies we think can successfully make IC. Our interest is in discovering what formulation of IC applies to sacred investments like those in works of art, animal species, human fetuses, and human lives. All these sacred values are sacred in one very special way that Dworkin has picked out: they are *genetically* sacred (LD 74-75). What makes these investments sacred is essentially connected to the history of the account, to the creative process embodied in each object. With genetically sacred objects, Dworkin tells us, people view the depositor's action—the creative process—as itself sanctifying the account. Because of this, it seems reasonable to expect that a fairly strict interpretation of IC would apply to all people's treatment of genetically sacred objects, irrespective of the particular way in which certain groups of people (for religious reasons) order natural and human investments compared to each other. We recognize that a strict interpretation is required because of a purely general feature of Dworkin's account: the quality of genetically sacred objects. Can we derive, from within Dworkin's own account, an appropriate interpretation of this (neutral) principle?

One plausible interpretation of IC regarding genetically sacred accounts might be derived from Dworkin's distinction between the two bases of the sacred—one broad source of creative investment is natural, the other human. Let us say therefore, as a theory-generated hypothesis, that IC may be violated for a genetically sacred account when a party makes what I shall call a *crossbasal* withdrawal from such an account. My suggestion is that many people feel that a party violates IC—and thus may illegitimately interfere with whatever the investment was toward—when that party makes a crossbasal withdrawal. If humans close a genetically sacred account that was opened and sustained by nature, for example, this is a crossbasal violation of IC. So formulated, IC may explain cases that are mysterious if one uses Dworkin's weight-based approach alone.

When a species of tree or animal naturally goes extinct, there seems to be a great subtraction of nature's creative investment. Yet people do not feel here (as they would in a case where the same destruction was caused by human actions) that such an event is a violation of the sacred. Dworkin's purely weight-based interpretation of the investment model cannot explain this important difference. But if we supplement Dworkin's account with a principle expressing concern for IC, the explanation of people's feelings in such cases becomes straightforward. Since nature (theistic/secular) opened the account embodied in the tree species, when nature closes the account, IC has been preserved. In the face of the epistemic block, the inability of humans to understand the terms imposed by nature's intentionality, this gives people confidence that the event fulfills rather than frustrates the investment.

Imagine a different case whereby a community of young lumberjacks had developed a (to them) richly rewarding and meaningful life project centrally committed to the project of efficiently cutting down trees. Unfortunately, the remaining stand of some (as it happens, uninteresting) tree species gets in their way and is destroyed. Since the lumberjacks' lives represent a creative human investment (they "live to cut") it may be a matter of calculative dispute as to whether the increase in the sacred investment that accrues in their lives could equal or surpass the frustration of natural investment deposited in this tree species.[8] But people who are concerned with sanctity—and perhaps especially people whose concern takes a liberal nature-admiring form—feel that in the lumberjack case a worse thing has happened. By Dworkin's weight-based interpretation of the investment model, this is mysterious: since what is destroyed is a product of strictly natural investment, we cannot explain their strong condemnatory view by the greater weight they give to human investment compared to natural investment. Indeed, that bias on their part may make them feel that the lumberjack case is less of a net cosmic shame. But they feel something terrible has happened no matter how their direct calculations about weight come out: a principle has been violated. The principle, I suggest, is IC and the violation is a crossbasal one. In both cases, the eradication of an entire tree species involves a great subtraction from what is wonderful in the world—in Dworkin's elegant phrase, a "cosmic shame" (for example, LD 75). But, faced with their uncertainty about what nature intends, a concern heightened when what is at stake is genetically sacred, people feel the loss in the lumberjack case is worse because of the way the loss comes about: that loss involves what we might call a *sacred violation*.

There is a sense in which people—especially people who put a high value on human investments—can recognize crossbasal violations in the opposite direction as well. Imagine a young doctor fully trained who is just starting out on her career, but is suddenly struck and killed by lightning. In such a case, we despair at the great waste of creative investment, but we can distinguish our feelings coming from that source from other feelings we have which in some ways are angrier. When a young doctor is suddenly taken from our midst by some freak accident—a so-called "act of God"—it is more than just the brute subtraction of creative value, even measured on a sophisticated frustration metric, that is upsetting. The lives of such people represent accounts not only of their own creative savings, but also a great investment of cultural and societal resources, a great human creative effort. IC has been violated when what happened was not just something impersonally to regret, but something that has an additional badness, a badness to which people are sensitive as investors. People may feel this extra badness without feeling in any way that their rights were violated.

---

[8] On one side of the balance is the intrinsic value of their life projects (aside from whatever personal value those projects may generate); on the other side is the intrinsic value of what is endangered by their life projects.

Recognizing IC as an essential component of the frustration metric would also enable us to explain people's views about active euthanasia which Dworkin finds mysterious. For example, Dworkin notices that the pattern of law that has emerged in the United States allows people to choose to die slow deaths by refusing life-saving/sustaining treatment while it forbids them to choose the quick deaths doctors could easily provide. Dworkin calls this "an apparently irrational result" (LD 184; also 254). But if we see that legal situation as reflecting people's widespread concern about crossbasal violations of IC, the mystery evaporates at once. People feel morally uneasy about humans stepping in to end a still-functioning biological life, even when the particular investments that can be tied to the intentionality of the person whose life it is already have been fulfilled substantially or as substantially as they are likely to be, and even when (through brain damage) the very possibility of human intentionality has been lost.[7]

---

[7] IC would make sense of the sentiment behind the Catholic Bishops's statement that "We believe that our following of Jesus Christ is most of all to die as he died—freely, with trust in God's plan and God's timetable, not ours!"—a sentiment left opaque by Dworkin's narrowly weight-based interpretation of the model—"Michigan Catholic Conference Statement on Physician-assisted Suicide and Euthanasia," *Origins*, XXIII, 4 (1993); LD 63-64. Further, IC gives a more precise sense to the concern Dworkin reports many people having that active euthanasia "*cheats* nature" (LD 214; emphasis mine). Of course, it may sometimes be that natural investments "run out," leaving only human ones (for example, a person conscious but on terminal life-support). Perhaps in other cases human investments "run out," leaving only natural ones (for example, a person demented but biologically self-sustaining). To me, this suggests that a concern for crossbasality, at least via a preliminary analysis, might generate a prohibition of active euthanasia, but not of passive euthanasia.

I am not suggesting that people think violations of IC are all or only crossbasal ones: sometimes IC can be violated in a way that is *intrabasal.* We think it a terrible thing, an act of desecration, when someone rushes into a museum and smashes a wonderful sculpture with a hammer. But we do not find it a shocking desecration when we learn that some painters paint over their own canvases, or that sculptors routinely pound out the very complex and beautiful clay mock-ups they make. Indeed, we find a certain poignancy in decisions like that of the American photographer Brett Weston who, on his eightieth birthday, burned the original negatives that had been his life's work, explaining that he had produced from them precisely the number and quality of prints he had intended. We may feel the acts of these "creative destroyers" are a great subtraction; but we do not feel here, as we feel with the museum vandal, that the acts are sacred violations. IC also helps explain the paradoxical attitude we have toward the art collector who destroys his purchases. He "has a right" to do with his property what he pleases, but he cannot wipe out the creative investment without (intrabasally) violating IC: such an act would be not merely a subtraction, but also a sacred violation—a violation that even his right cannot disguise. I believe there are many further patterns of intrabasal violations (some, like these, *crosspersonal*) that might be charted as underlying people's beliefs regarding the treatment of things with intrinsic value,[8] though it is the simple and more pervasive aversion to crossbasal violations which concerns us.

There seem to be at least two further elements of this important 'Who does it?' question. First, when people ask the 'who' question, they also seem interested in the manner in which events involving intrinsically valuable objects occur. For example, we feel differently about the intrabasal violation when we learn that the museum statue was smashed, not intentionally, but by the janitor's accidentally bumping it over with his mop. While intentionality (or something like it, see below) always seems necessary on the input side to make creative contributions investments, the role of intentionality on the withdrawal side—what we might call the *how* element—builds out in complex ways from the system of intrabasal and crossbasal distinctions we have already charted.

---

[8] Indeed, I suspect that a uniquely appropriate interpretation of IC could be worked out for each species of sacred value that Dworkin describes. Consider associationally sacred things. While many Americans were appalled in the late 1950s when young Chinese people started burning American flags to protest United States government policy, I submit that people felt a qualitatively different outrage when in the 1960s young *Americans* started treating American flags the same way. With things sacred by association (quite unlike with genetically sacred things) destructive acts are often judged worse when carried out by people most closely associated with those things.

Whether it is nature or humans acting to close an account, people's impersonal evaluations of events involving intrinsic value seem sensitive to a further set of questions corresponding, roughly, to categories of *destruction by chance* versus *destruction by nonchance*. It is difficult to discern the precise patterns that such sensitivity takes for most people—if indeed there are precise patterns here at all. Certainly, some human crossbasal violations that occur through chance (that is, unintentionally) are evaluated less harshly than human crossbasal actions that happen through nonchance (intentionally): imagine, for example, that the remaining stand of the tree species is crossbasally destroyed not by the lumber jacks' saws but by one of them accidentally dropping a match.[9] With acts of natural destruction, by contrast, the chance/nonchance distinction is probably best understood as a distinction between destruction through chance and destruction as a predictable result.[10] For example, from the perspective of an intrabasal concern for intrinsic value, people may feel differently about the destruction of the Tyrannosaurus Rex when they are told that the species went extinct not as an inexorable result of competition with better designed creatures (that is, predictable result) but because of the sudden environmental changes that occurred when the earth was hit by a stray meteor (chance). If people are sensitive to a difference here perhaps it is because, while they think animal species are wondrous precisely because of their evolved traits, the effects of the stray meteor seem somehow outside the predictable environmental parameters of competition and selection in which those admired traits typically arise.[11]

There is a further element of the 'who' question we have been considering: for when people care how an act is done (whether by chance or nonchance), they also care *why* it is done. The death of a young doctor by lightning seems less bad than the murder of a doctor for fun—though both involve crossbasal destructions.[12] A natural flood that kills people, while tragic, seems less bad than a deadly flood caused by malicious human sabotage. Indeed, the importance people attach to the quality of agents' motives might be thought to generate an objection to the entire account of IC which I have given so far: perhaps, in all these cases, what people find bad is simply the human choice for evil and not the crossbasality at all.

---

[9] As with the clumsy janitor, in such cases questions of negligence may push our evaluations from the category of chance toward that of nonchance—yet there remains conceptual space for the genuinely accidental.

[10] I thank Bernard Williams for suggesting this terminology to me.

[11] In many cases, especially with crossbasal violations that occur in the nature-against-human direction, the question of intentionality threatens to confound the very categories "natural" and "human": with the young doctor killed by an "act of God," for example, would it make a difference if she had been struck by a bus rather than by lightning?

It is easy to misunderstand the force of this objection. Our strong reactions to murder for fun or malicious sabotage are bound up with our abhorrence of the rights violations in those examples. We must try to peel away our reactions about people's interests, since our concern is with intrinsic (not personal) value. One way to do that is to reconsider crossbasal cases, such as the human destruction of plant and animal species, which do not violate the rights of humans but seem like choices for evil nonetheless. But when we ask again how we know that the human destruction of the snail darter or the African elephant is worse than its natural destruction, then the suggestion itself (that people really care about the quality of the choice, not the crossbasality) becomes unclear. For on the account we have been developing, strictly, people care ultimately not about the crossbasality but about the value regarding which the crossbasality is an epistemic indicator: people care that creative investments be fulfilled rather than frustrated. Faced with the epistemic block, IC gives people guidance about how they and others concerned about that value should behave.

This in mind, there are various reformulations of this objection which should be guarded against. For example, instead of saying that people care not about crossbasality but about the quality of the choice itself, it would make little sense to suggest that people care not about the underlying value (that is, respecting investments) but about the quality of the choice. Since people's evaluations of the quality of the choice must proceed in terms of some criteria (to distinguish evil choices from good), on the sanctity model those criteria must come from concerns about how the choices affect the value. Nor would it help, alternatively, to explain such cases by saying that the value somehow consists in the choosing, for it is basic to the frustration metric on which the sanctity paradigm is founded that nature's contributions as much as human ones be thought of as undergirded by intentionality or something like it. Since both are investments, there is "choice" on both sides. If someone—say an extreme liberal—were to insist that sanctity consists only in the designs and choices of humans (whether as depositors or withdrawers), that person would have abandoned Dworkin's paradigm altogether, just as much as an extreme conservative who insists that sanctity consists only in the designs of God/nature.

---

[12] The destruction by lightning represents a violation by nature against the human investment in the doctor's life; murder is a violation by humans against the natural investment in that life.

It is tempting to say that most people (liberal or conservative) think the murder of the doctor is worse than her death by lightning simply because in the former case (but not the latter) the destroyer was in no part an investor in what is destroyed—as though the 'why' question could be completely reduced to the main crossbasal aspect of the 'who' question. This reduction is probably too simple. What if Weston had destroyed his wonderful negatives out of a desire to inflate the value of his existing prints (that is, a concern for personal, instrumental value) or out of dementia (that is, for no value at all)? From the perspective of a concern for intrinsic value, and despite the power of the reduction I mentioned, people do seem to have a residual concern for the quality of the motive as a freestanding epistemic indicator. If people choose to destroy, we feel easier if they destroy (in some sense) as creators.

Much more could be said about the *how* and *why* aspects of IC.[13] It is enough here to see that, within the investment paradigm, a concern for the quality of people's choices does not compete directly with a concern for crossbasality, as though either were an ultimate value. Whether the two can be said to compete for primacy as epistemic indicators about what respect for creative investments actually requires is a different (and more difficult) question. We shall return to that question when considering the liberal position on abortion.

Dworkin might well be criticized for offering little empirical support for his substantive claims about people's attitudes toward things they treat as intrinsically valuable. In arguing for the importance and pervasiveness of people's concerns about crossbasal violations of IC, I have been making empirically unsubstantiated claims of just the same sort. While I, like Dworkin, have been relying on what I hope is the uncontroversial nature of my main substantive claims, it is worth noting that Dworkin at least would be unlikely to disagree with me about the importance of IC regarding genetically sacred things. For while Dworkin seems quite unaware of this crucial aspect of the sanctity model, his own examples, again and again, implicitly rely on people's concern about crossbasal violations of IC. For example, in describing the badness people feel about the creative waste when a species of animal goes extinct, he talks not about how badly people would feel if the Asian tiger were to go extinct as part of the natural cycle of biodiversification and contraction, but about how badly people would feel "if we [humans] stamped out the Asian tiger [nature]" (LD 80).[14] Of course, people may often think that any creative subtraction is regrettable (an animal species dying out naturally; the sculptor pounding out his mock-ups); but it is striking that in seeking to make vivid people's regret about waste, Dworkin repeatedly chooses examples that get their special potency—just as I claim, and as the epistemic problem would have us predict—by involving a crossbasal violation of IC as well. Anyone who takes Dworkin's investment model seriously seems likely to find himself, whether self-consciously or not, expressing a concern for IC.

---

[13] For example, there are complex asymmetries regarding both the ways humans *ascribe* intentionality to nature versus to humans, and the ways humans *evaluate* that intentionality once ascribed.

IV. ABORTION AND RESPECT RECONSIDERED

What if I am right? What if Dworkin has missed an important series of questions about this new model he is advocating? What if, by asking those questions, we discover a component—IC—which is required to make epistemic sense of the judgments about frustration and fulfillment which Dworkin sees as the model's operative core, a component which when recognized offers ready explanations for cases that Dworkin admits are puzzling or irrational by his own less-developed account of the model?

It is the central conclusion of Dworkin's moral argument that the new abortion model he proposes—the investment model—leaves it an open question as to whether or not a decision to abort best respects the intrinsic value of life. What is more, it is crucial to his argument that this question be left open in a very specific way: the morality of such decisions must be seen to turn directly on calculations about the relative weights people assign to natural or human investments. Since that direct assignment of weights occurs incrementally, with reference to people's deepest beliefs about the meaning and importance of human life, the model never generates a conclusive external answer to the question of whether or not abortion is immoral. Liberals give greater weight to the human contributions represented by the intrinsic worth of the life projects of the pregnant woman; conservatives give more weight to nature's contribution represented by the intrinsic worth of the fetus. According to the sanctity model, says Dworkin, the public disagreement is actually just a matter of degree, a gradated dispute over the proper assignment of relative weights.

---

[14] There are many examples of Dworkin's unwitting reliance on people's special concern about avoiding crossbasal violations of the investment connection regarding genetically sacred things (see, for example, LD 75, 76, 79).

When one sees the whole of Dworkin's model, as I have tried to reveal it, these conclusions become unsteady at best. If, as Dworkin says, an early-stage fetus has a basis in natural investment, then any human decision to kill it, to close the account, is prima facie a crossbasal violation of IC.[15] People in a position of epistemic uncertainty must consider this. This means that when Dworkin encourages people, both liberal and conservative, to abandon the old rights-based paradigm and to rethink their views about the morality of abortion via the investment model, both groups will almost certainly disappoint him. For neither group will think the thoughts (or reach the conclusions) that Dworkin expects.

Consider first the conservatives. To adopt the sanctity model, Dworkin says conservatives must abandon the justice-based paradigm, and thus give up their hallmark claim that each and every fetus has a right to life. Dworkin says the model allows conservatives to maintain their moral opposition to abortion, however, since it allows them to claim that, because of their particular views about the meaning and point of human life, natural investments always count for more than human ones, and thus abortion is always a greater net waste than conception.

Dworkin has seriously underestimated the power of his own model, however, when it comes to capturing the conservative view. Conservatives, by the terms of the sanctity model, see abortion as an especially objectionable crossbasal violation of IC: an arrogant human decision to close an account miraculously opened by God. "Man must not undo what God has wrought" is how they put the point of IC in the terms of their own religious system. When its principled IC dimension is made explicit, the sanctity model captures the very essence of conservative opposition to abortion. For conservatives, abortion is not merely "a cosmic loss," a "bad thing to have happened," or "something impersonally to regret"—though abortion is bad for those reasons, too. The moral issue is hotter than that and the investment model, when developed, can explain precisely where the issue gets its heat. The investment model tells conservatives that abortion is a moral *transgression*, a sacred *violation*, an act proscribed *by principle*—just as they claimed all along.

---

[15] "Any human creature, including the most immature fetus, is a triumph of divine or evolutionary creation and of the process of human reproduction that produces something so fantastic from, as it were, nothing. The human investment begins *almost* as soon, and steadily increases throughout a normal life" (LD 83; emphasis mine).

What about the liberals? For liberals to adopt the sanctity paradigm, Dworkin says they must consider not only their special concern for human investments like those represented by the life projects of pregnant women, but also make explicit the awe they feel for the creative process at work in even the earliest-stage fetus.[16] Making this request of liberals is far less innocuous than it may appear. Unfortunately, Dworkin's only comments are oblique on the question of which base, natural or human, an early-stage fetus more properly represents.[17] Perhaps Dworkin recognizes that his crucial distinction between human and natural bases of investment takes a particularly puzzling form in the case of human reproduction (since, on his account, the biological aspect of human reproduction is "natural" not "human"). Still, in the case of a genetically sacred object such as an early-stage fetus—a thing we respect because of our special reverence for the process that produced it—the problem of determining the main contributory base (human or natural) seems surmountable. In the case of each early-stage human fetus, which creative process most inspires our awe: (1) the human one—the act of sexual intercourse, and the subsequent dietary and health decisions of the pregnant woman—or (2) the natural one—the mysterious process of selection and genetic individuation (or creation) that gave rise to this maturing biological (spiritual) organism? When we honestly ask ourselves which process makes every human fetus awesome, even for extreme liberals (who have the highest regard for human accomplishments) the answer seems clear. Even when conception is maximally human—planned and intended as an expression of a couple's real love for each other—what is created (the developing early-stage fetus) seems most awesome because of what nature contributes, for the glimpse each growing fetus provides of a complex creative process quite beyond any merely human craft.[18]

---

[16] Dworkin says this will allow liberals to make sense of their sentiment that abortion is a morally weighty matter (more properly a cause for regret than, say, an appendectomy), even while maintaining their conclusion that—because of their particular view about the meaning and point of human life—human investments count for more than natural ones, and thus a decision for abortion may involve less net waste.

[17] Dworkin sees early fetal life always as representing significant natural investment from the moment of genetic individuation. At one place, Dworkin suggests that human investment is present that early (only?) in cases where reproduction is planned (LD 83). More often, however, Dworkin considers the human investment in an early-stage fetus "negligible" (LD 99).

The problem is that once liberals do what Dworkin asks and recognize the sanctity of even the earliest-stage fetus, they immediately become vulnerable to a challenge from those conservatives with whom they share the sanctity model, a challenge delivered in an unexpected form. For Dworkin's model allows conservatives to explain in definite terms why abortion is wrong, even setting aside their own particular beliefs that natural/divine investments are weightier than human ones (beliefs which, to them, may make abortions seem "especially objectionable" or "arrogant"). If a fetus has a creative basis in nature, conservatives using Dworkin's model might point out, then, in the face of the epistemic block, a human decision to kill it is forbidden by the principle at the frustration metric's operative core: abortion is wrong, a crossbasal violation of IC.[19] Quite contrary to Dworkin's expectations, a religiously neutral objection to abortion is built into the very base of the investment model.

At that point, the best recourse for liberals committed to the sanctity model might be to deny the importance of the crossbasal aspect of IC as a solution to the epistemic block. Even if a concern for crossbasality is religiously neutral (since recognizing its authority does not rely on tendentious claims about natural versus human investment weight), still liberals may say that crossbasality—with its rigid concern for *who* over *why*—remains in a deeper sense a typically conservative principle, founded more on superstition than reason, designed to limit human action and decision making out of a purported concern for what is "natural." Because of their own very different understanding about the place of human life in the universe, liberals may say that their concern for sanctity requires that they always give epistemic primacy to their evaluations of the quality of people's choices. People show respect not when they slavishly follow some rule, but when each makes her own decision with seriousness and care. With abortion especially, the conscience of each person is the best moral guide.

But this new pattern of dispute is beginning to sound depressingly familiar. Conservatives passionately affirm a choice-limiting principle that liberals passionately reject: there is no conceptual middle ground on which a compromise could be struck. People who use the new sanctity model—conservative and liberal alike—find it directing them back behind barricades just as stark as the ones they manned when the debate was cast in the old justice-based way.[30] Dworkin's proposal for a new way forward, when followed a few steps further, turns out to be just another path back.

---

[18] This also makes sense of the feminist description of unwanted fetal development as something happening *to* women rather than *by* them.

[19] By highlighting the 'who does what?' questions, IC could also help generate more detailed explanations for the various exceptions conservatives allow (see footnote 4 above).

This would be an unhappy enough conclusion for the sanctity model, but for liberals the moral matter looks likely to get even worse. Even if the investment model directs both sides back behind their familiar barricades, the model ensures that the liberals now will be much less comfortable defending their old moral position. No doubt mainstream liberals would like to say that they always give epistemic primacy to conscientious 'why' questions over 'who' ones. But that seems untrue. Indeed, when it comes to crossbasal violations of other sorts of natural investments—regarding snail darters, or certain species of owl—liberals tend to be especially concerned to maintain the primacy of the 'who' question. These liberal attitudes are just what the foundations of the investment model require, regarding human choices about destruction of natural investments in species of animals and plants, and regarding their choices about the destruction of human fetuses as well. After all, what would it mean, within the terms of the investment model, for a human "conscientiously" to choose abortion? Presumably, such a choice would be "conscientious" if it were the output of a fair weighing of the conflicting investments, the investment in the woman's projects versus the investment in the early fetus, a calculation that requires an understanding of the towardness of each investment. But since the early-stage fetus is a product of natural investment, this is precisely the kind of case where the epistemic block presents itself most darkly. In the terms of Dworkin's investment model, how else could one evaluate the quality of people's choices except by considering the crossbasal aspect of IC?

Of course, liberals committed to the moral permissibility of abortion could reject the entire investment paradigm—leaving it to the conservatives. But our aim throughout has been to find out about abortion by considering the patterns underlying people's whole range of moral attitudes, rather than to let their position on abortion, however emotional and entrenched, somehow determine all the rest. If people on both sides of the abortion controversy, liberals as well as conservatives, become open minded enough genuinely to reconsider even the basic groundings for their position, it is the liberals who will be most discomfited by Dworkin's new paradigm.[n]

---

[30] Notice that Laurence Tribe, for one, has picked out Dworkin's hope of finding a new pathway to accommodation as the main ambition of Dworkin's abortion project—"On the Edges of Life and Death," *The New York Times Book Review* (May 16, 1993), p. 41.

V. LIBERALISM AND THE PROTECTION OF SANCTITY

The legal implications of Dworkin's investment model are complex and, because of the unsettled nature of constitutional argument about abortion, the stakes are high. If one thinks liberalism is only about the protection of individual rights, then there is no problematic anti-abortion argument to be raised here: the interest of Dworkin's proposal will end with the moral challenge it puts to liberals. But Dworkin himself believes that the liberal state can have a responsibility to act not only in order to protect rights—what he calls the state's *derivative* responsibility—but also in order to protect intrinsic values—its *detached* responsibility (UR 396-98). Dworkin says we must attribute detached obligations to the liberal state so that we can explain why such a state may sometimes limit liberty when seeking to protect intrinsic values like those represented in great works of art, animal species, traditions of culture and craft, historic buildings, and whole future generations of people (for example, LD 154).

Of course, as Dworkin notes, if one thinks the liberal state properly has such detached responsibilities, and one also insists that most people (conservative and liberal) accord intrinsic worth to early fetuses, then the investment model would appear to provide a new way of structuring the legal argument of extreme conservatives who think abortion should be outlawed. If extreme conservatives allow their political thinking to be structured by the investment model, they will argue that since the liberal state can limit the liberty of people whose actions threaten animal species or historic buildings, it must have a detached responsibility of just the same sort to limit the liberty of those who would destroy an even more important intrinsic value—the value of each developing fetus.

---

[n] IC will challenge conservatives, too—for example, asking some conservatives to reexamine the exception they allow themselves regarding human violations of IC when it comes to nature's investments in species of flora and fauna (of course, mainstream conservatives could reply that they are as opposed as anyone to the actual destruction of species entailed by such crossbasal violations; they simply judge the risk differently). Further, it will require conservatives to provide a more detailed explanation of their advocacy of capital punishment, an apparent state-based crossbasal violation of IC (though capital punishment for murder might be justified within their system as an in-kind response to a previous crossbasal violation). More substantively, a concern for crossbasality would probably ask conservatives to reconsider their moral opposition to passive euthanasia (cf. footnote 7 above).

Dworkin is well aware that the investment model channels the extreme conservative's legal argument in this new direction. Because of the narrowly weight-based way he understands the investment model, he welcomes this new pattern of argument—believing that he can contain it. Generally, Dworkin believes that the acceptance of the sanctity model on the moral level may make parties on both sides more willing to seek accommodation on the political level, since (on Dworkin's reading) the sanctity model transforms the moral dispute about abortion into a gradated one that turns ultimately on a respectable difference of opinion regarding the relative assignment of weights of human and natural investments. In hopes of showing how the United States Constitution can provide that sort of political accommodation, Dworkin offers a three-point privacy test to protect a woman's right to seek an abortion without eroding the state's detached responsibility to limit liberties that threaten other intrinsic values: "A state may not curtail liberty, in order to protect an intrinsic value, when [1] the effect on one group of citizens would be special and grave, when [2] the community is seriously divided about what respect for that value requires, and when [3] people's opinions about the nature of that value reflect essentially religious convictions that are fundamental to moral personality" (LD 157; bracketed numbers mine).[22] Dworkin calls this the *principle of procreative autonomy* and finds a primary textual home for it in the freedom of religion clauses of the First Amendment (UR 418-20).

As we have seen, however, the investment model gives no reason for parties to be more open to accommodation on the question of abortion. Dworkin's new model not only fixes both sides behind their old moral barricades, but does so in a way that energizes the conservatives. Because of this, Dworkin's three-point privacy test would come under a great deal of pressure. Extreme conservatives might point out that, if this test is allowed to block the state's detached responsibility to protect fetal life, then the same test could be used to prevent states from limiting liberty in a great range of other cases, including many cases dear to the liberal side. For example, Catholics might use Dworkin's test to argue that the state cannot tax them to support exhibits containing certain works of art they find objectionable. Commercial logging interests might use it to overturn environmental regulations that threaten the lifestyles and folkways of their workers. Because Dworkin sees intrinsic value in so many different kinds of processes, activities, and things—insisting, for example, that every person's life has intrinsic value in virtue of that person's creative investment in his own particular projects and commitments—and because of Dworkin's expansive definition of "essentially religious" beliefs,[23] the privacy test is far too strong. Amid the other assumptions associated with Dworkin's sanctity model, it amounts to a strong and general principle of autonomy against a value-protecting state.[24] (There is a Scylla with this Charybdis: for if a test like Dworkin's were made weaker, it might not generate a secure right of abortion against a majority passionately pursuing (its view of) the state's detached obligation to protect life.)

---

[22] For Dworkin, beliefs are "essentially religious" (in the constitutionally protected sense) if they involve people's deepest views about the meaning and importance of human life and the place of humans in the universe (UR 422-24). Apparently, Dworkin sees "essentially religious" beliefs as a subset of people's beliefs about intrinsic values (see, for example, LD 165).

Dworkin insists that a liberal state can sometimes limit liberty to protect intrinsic values. Further, he would like to see the public debate over intrinsic values restructured along the lines of his investment model. Given that assumption and that objective, Dworkin needs some more adequate legal test that explains when the state can act to protect such values and when it cannot. Such a test—if we could find one—should meet the requirements of law's integrity: as Dworkin says, "a judge who adopts a principle must give full weight to that principle in other cases he decides or endorses" (UR 394).[25]

---

[23] On Dworkin's expansive conception, atheists have essentially religious beliefs (UR 413). T.M. Scanlon suggests, tellingly, that Dworkin's notion of the religious may be so wide as to "apply as well to conceptions of justice and to other moral ideas which the state clearly can enforce"—see "Partisan for Life," *New York Review of Books* (July 1993), p. 47.

[24] Briefly, the first two points of Dworkin's test fairly obviously could not distinguish the dispute about abortion from disputes about other intrinsic values (for example, regarding religiously offensive art, or the proper limits of environmental protection). But the third point, when considered amid Dworkin's other claims concerning the sanctity model, is also unhelpful. First, Dworkin sees intrinsic value inhering in the life projects in every person's (and community's) projects and commitments; these values (beyond their personal value) require consideration by any state acting to protect even those intrinsic values which those projects threaten. From the other side, a central part of one's moral personality is defined by the way one understands the relation of his own actions (including the intrinsic worth of his own projects) toward other intrinsically valuable objects out there in the world: on the terms of Dworkin's model, the woodcutter, as much as the pregnant woman, is making a personality-defining choice (a choice which engages his deepest beliefs about the role of human existence within the universe) when he balances the intrinsic worth of his own projects against the cosmic value of those other parts of creation his projects threaten.

[25] "The law's integrity demands that the principles necessary to support an authoritative set of judicial decisions must be accepted in other contexts as well" (LD 158). See also *Law's Empire* (Cambridge: Harvard, 1986), chapter 7.

So, for example, this test should be able to generate decisions compatible with existing decisions that protect endangered species from destructive human projects—no matter how important those projects may be to particular people or their communities. A commitment to integrity would also lead us to hope for a test that would support established judicial decisions about the limitations of liberty regarding the taking of human life in other contexts—for example, legal rulings regarding human decisions to cause death at the other end of life, decisions like those involved with proposals for active euthanasia.

It is difficult to make out such a test in a way that does not sound too unfamiliar, at least in part because Dworkin's claim that the liberal state has detached responsibilities to protect such values is itself so unfamiliar. Still, the legal adoption of the model Dworkin proposes would urgently require that we construct some such test. If Dworkin's three-point privacy test proves unable (in Dworkin's sanctity-laden world) to make the distinctions that legal integrity requires, perhaps one should look for a better test in the core of Dworkin's own model: if a liberal state can ever limit human liberty to protect intrinsic value, it can do so to prevent crossbasal violations of IC.[26]

A legal test based on a concern for IC would firmly cohere with, and help better explain, the dominant pattern of American justification for decisions upholding prohibitions on certain types of euthanasia. Even when a person's life seems finished from the perspective of normal human projects, respect for nature (and the possibilities for that life which we cannot foresee or comprehend) requires that no human hand be allowed to end that life's unfinished biological story.[27] Further, unlike Dworkin's test, a legal test based on the crossbasal aspect of IC would reliably support the American legal orthodoxy in the main run of environmental cases.

---

[26] Of course, legal integrity would require different interpretations of this test in different countries. In Ireland, where the state sometimes claims detached obligations to limit liberty to protect Catholic sacraments, a concern for IC might even justify the proscription of legal divorce (giving content to the ceremonial imperative, "What God has joined let no man tear asunder")! In the United States, legal integrity would require a more moderate—though still worrying—interpretation of the IC test (see immediately below).

[27] Thus, a concern for IC might provide a powerful explanation of Chief Justice William Rehnquist's claim in *Cruzan v. Director, Missouri Dept. of Health* about Missouri's "interest in the protection and preservation of human life," an interest abstractable from the interests of the young woman living that life (497 U.S. 261 (1990), at 280).

By the IC test, for example, even the great impersonal importance of human projects cannot justify the deliberate (crossbasal) eradication of one of nature's species: the liberal state—to protect an intrinsic value—could justifiably stop the lumberjacks from cutting down their endangered trees.[38] Of course, the worry for people in favor of abortion rights, liberals and moderate conservatives alike, is that the same test would simultaneously provide a new legal explanation of why a liberal state can justifiably block pregnant women from aborting their growing fetuses. For in such cases, the argument would go—no matter how intrinsically weighty the human projects on the other side, and even granting *arguendo* that the early-stage fetus as yet has no interests in its own life—still the liberal concern for sanctity, our shared sense of awe at the complex natural miracle unfolding in each growing biological life, would require that human acts of fetal destruction be legally proscribed.

JOHN TOMASI

Brown University

---

[38] Such a test might help explain the Supreme Court's underlying reasoning in *Tennessee Valley Authority v. Hill*, which prevented an enormous human project—the Tellico Dam—from threatening a minor species of fish (437 U.S. 153 (1978), especially at 169, 184-87). (Though I here leave aside the complex question of whether the subsequent legislative maneuvers by Congress to get around that ruling did anything to undermine that reading of the judicial precedent.)

VAGUE IDENTITY AND FUZZY LOGIC*

In a five-line derivation, Gareth Evans[1] reduced to absurdity the assumption that some identity—any identity—is indeterminate, or so he claimed. His notation "allow[s] for the expression of the idea of indeterminacy by the sentential operator '∇' " (*ibid.*, p. 208). The derivation runs as follows:

(1) $\nabla(a = b)$
(2) $\lambda x[\nabla(x = a)]b$ from (1)
(3) $\sim\nabla(a = a)$ axiom
(4) $\sim\lambda x[\nabla(x = a)]a$ from (3)
(5) $a \neq b$ from (2) and (4) by Leibniz's Law

Evans remarked that (5) "contradict[s] the assumption, with which we began, that the identity statement '$a = b$' is of indeterminate truth value" (*ibid.*).

The reason for employing Alonzo Church's lambda notation for property abstraction (whereby the statement 'a has the property F' is represented '$\lambda x[Fx]a$') is brought out by the following simplified version of the derivation:

(1a) $\nabla(a = b)$
(2a) $\sim\nabla(a = a)$ axiom
(3a) $a = b \rightarrow (\sim\nabla(a = a) \rightarrow \sim\nabla(a = b))$ from Leibniz's Law
(4a) $a \neq b$ from (1a), (2a) and (3a) by propositional calculus

Line (3a) is an obvious weak point, for it is well known that in contexts containing (what Arthur Prior called) quasimodal operators, of which $\sim\nabla$ is an example, the substitution of identicals can turn a truth into a falsehood. Much more compelling is the analagous form of (3a) in which property abstraction is employed (assuming, of course, that one is prepared to countenance *being indeterminately identical to a* as a genuine property):

(3b) $a = b \rightarrow (\sim\lambda x[\nabla(x = a)]a \rightarrow \sim\lambda x[\nabla(x = a)]b)$

In association with lines (2) and (4) of the original derivation, (3b) yields the desired conclusion.

---

* I would like to thank Ross Brady, Diane Proudfoot, and Tim Williamson for helpful discussions and correspondence.

[1] "Can There Be Vague Objects?" *Analysis*, xxxviii (1978): 208.

0022-362X/97/9410/514–34

Notoriously, Evans's brief paper left unclear the precise moral he wished to draw from this supposed reductio. Although he did not say so in the paper, his target was in fact Michael Dummett's view that the world *in itself* might be vague. Evans makes this clear in a letter written to David Lewis in November 1978:

> The author I had in mind was Dummett, who has written and spoken of vagueness being in the world, and so on, but I did not refer to it because then as now I cannot find a reference in his published writings.[2]

Perhaps this passage from Dummett's *The Interpretation of Frege's Philosophy*,[3] in which he discusses his "Wang's Paradox,"[4] would have served Evans's purpose:

> I [said] in "Wang's Paradox" [that] "the notion that things might actually *be* vague, as well as being vaguely described, is not properly intelligible" [p. 314]. I should now like to withdraw this remark.... [I]t is natural to us to conceive of physical reality as, in itself, capable of description in absolutely precise mathematical terms.... This inclination does not, however, appear to be more than a prejudice.... It is not apparently absurd to suppose the contrary, namely that the physical world is in itself such that the most precise description of it that even omniscience would yield might yet involve the use of expressions having some degree of vagueness (*op. cit.*, p. 440).

In the opening sentences of his paper, Evans gives a tolerably clear statement of the Dummettian position that he wishes to challenge:

> It is sometimes said that the world might itself *be* vague. Rather than vagueness being a deficiency in our mode of describing the world, it would then be a necessary feature of any true description of it (*op. cit.*, p. 208).

Regrettably, Evans gives his readers no detailed guidance as to how his derivation is intended to bear on this position.[5] Leaving aside the dark issue of vagueness-in-the-world, what can be said is that Evans's derivation presents a prima facie challenge to the view that '$\nabla(a = b)$' is true. Indeed, his paper is not uncommonly regarded as presenting a major objection—perhaps the principal objection—to the view that there are vague or fuzzy or indeterminate identities.[6]

---

[2] I am grateful to Antonia Phillips and to the Gareth Evans Memorial Trust for permission to quote from this letter, and to David Lewis for information and assistance.

[3] London: Duckworth, 1981.

[4] *Synthese*, XXX (1975): 301-24.

[5] Exegesis is offered by J.A. Burgess, "Vague Identity: Evans Misrepresented," *Analysis*, XLIX (1989): 112-19; B. Garrett, "Vagueness and Identity," *Analysis*, XLVIII (1988): 130-34, "Vague Identity and Vague Objects," *Noûs*, XXV (1991): 341-51, and "Vagueness, Identity, and the World," *Logique et Analyse*, CXXXV-CXXXVI (1991): 349-58; Lewis, "Vague Identity: Evans Misunderstood," *Analysis*, XLVIII (1988): 128-30; H.W. Noonan, "Indefinite Identity: A Reply to Broome," *Analysis*, XLIV (1984): 117-21, and "Vague Identity Yet Again," *Analysis*, L (1990): 157-62; D.E. Over, "Vague Objects and Identity," *Analysis*, XLIX (1989): 97-99; F.J. Pelletier, "Another Argument against Vague Objects," this JOURNAL, LXXXVI, 9 (September 1989): 481-92.

Evans's derivation falls short of being a reductio, for there is no explicit contradiction between (1) and (5). Evans seemed to think that the derivation can be extended to yield an explicit contradiction. He wrote:

> If 'Indefinitely' and its dual, 'Definitely' ('Δ'), generate a modal logic as strong as S5, (1)–(4) and, presumably, Leibniz's Law, may each be strengthened with a 'Definitely' prefix, enabling us to derive:
>
> (5′) $\Delta \sim(a = b)$ (*op. cit.*, p. 208)

It has often been remarked that the reasoning here is hard to reconstruct, for if '∇' and 'Δ' are duals, then '$\Delta(a = b)$' says that '$a \neq b$' is not indeterminate, and so is true if '$a=b$' is determinately false. (A pair of quasimodal operators $\uparrow$ and $\downarrow$ are *duals* just in case each is definable in terms of the other in the fashion of $\Box$ and $\Diamond$, viz $\uparrow A \equiv {\sim}\downarrow{\sim}A$ and $\downarrow A \equiv {\sim}\uparrow{\sim}A$.) Yet if the logic of '$\Delta$' is 'as strong as S5', then '$\Delta(a = b)$' implies '$a = b$'.[7] (In the letter previously mentioned, Evans himself agrees that the falsity of 'P' suffices for the truth of 'ΔP'.)

This slip of Evans's does not matter much, however, for even in the absence of an explicit contradiction, the derivation (1)–(5) is trouble enough for the view that there may be indeterminate identities. I shall use '$\vdash$' to represent the relation of deductive consequence associated with the proof. If the proof is not fallacious, we have:

$$\nabla(a = b) \vdash a \neq b$$

Contraposing and eliminating the double negation gives:

---

[6] For example, by J. Broome, "Indefiniteness in Identity," *Analysis*, XLIV (1984): 6-12, here p. 7; J.A. Burgess, "Vague Objects and Indefinite Identity," *Philosophical Studies*, LIX (1990): 263-87, here p. 268; P.F. Gibbins, "The Strange Modal Logic of Indeterminacy," *Logique et Analyse*, C (1982): 443-46, here p. 446; B. Johnsen, "Is Vague Identity Incoherent?" *Analysis*, XLIX (1989): 103-12, here p. 106; E.J. Lowe, "Vague Identity and Quantum Indeterminacy," *Analysis*, LIV (1994): 110-14, here p. 111; T. Parsons, "Entities without Identity," *Philosophical Perspectives*, I (1987): 1-19, here p. 12; M. Tye, "Vague Objects," *Mind*, XCIX (1990): 535-57, here p. 555.

[7] See also notes 31 and 44.

$a = b \vdash {\sim}\nabla(a = b)$[8]

The bearing of a result such as this concerning identity *statements* on the Dummettian claim that there might be vagueness in the world itself is far from clear. Nevertheless, this result, if correct, is a substantial one in its own right. Assuming that we are prepared to apply detachment to what we assert, the result tells us that vague identity statements are assertible only on pain of inconsistency: if it is the case that $\nabla(a = b)$ then asserting ‘$a = b$’ leads, via detachment, to the contradiction $\nabla(a = b) \& {\sim}\nabla(a = b)$. (Or if the principle derived in note 8 is employed, then asserting ‘$a = b$’ when the latter is not determinate leads to the contradiction: $\Delta(a = b) \& {\sim}\Delta(a = b)$.) This is just the sort of result which an opponent of vague identity needs, for it drives a wedge between identity and concepts that are uncontroversially vague, such as *bald* and *tall* and *blue*. As an admirer of the vagueness of ordinary language might put it, the practical value of a vague language is that speakers can assert what is *true enough* for their purposes. If Evans's result is correct, it shows that in the case of identity, nothing less than determinate truth can be truth enough. But is the result correct?

Fuzzy logic is widely canvassed as the logic of vagueness. Yet no one, I think, has explored the question of whether or not Evans's derivation can be carried through in fuzzy logic. I shall argue that it cannot.[9] (Several articles have investigated the fate of Evans's derivation in a three-valued setting, but three-valued logic is not fuzzy logic.[10])

Fuzzy logic aims to extend ordinary deductive methods to situations in which the information available may be only partly or approximately true. In fuzzy logic, each statement has some numerical *degree* of truth or falsity. These degrees are the truth values of the logic, and for full generality continuum many values are usually permitted. The three-valued approach, with its one grudging nonclassical value, has little to recommend it in the context of vagueness. With only three values to hand, all statements that are neither completely true nor completely false have to be thrown indiscriminately into the "neither" box. This wastes much significant information: for example, that *A* is truer than *B*, or that *C* is nearly but not completely true, or that drawing a given inference will produce a conclusion that is no less true than the weakest premise. The three-valued approach cannot generate a useable extension of the classical theory of inference, for in practical terms the effect of assigning a statement the catch-all value "neither" is simply to prevent us from usefully employing it as a premise in inference. It is of no practical use whatever to be told (as Jan Lukasiewicz[11] tells us) that, if either one or the other premise of an application of modus ponens has the value "neither," then the conclusion is either "true" or "neither." Whereas the information that, if both premises of an application of modus ponens are approximately true, then so is the conclusion, is exactly the sort of thing one needs to know in order to be able to reason with vague statements.

---

[8] The validity of contraposition in contexts involving ‘∇’ has been questioned (for example, by Parsons, pp. 9-10). Indeed, Evans's whole derivation has a contrapositive flavor. A more direct version, free of contrapositions, leads to the pleasingly blunt $a = b \vdash \Delta(a = b)$. The derivation is: (1) $a = b$ (assumption); (2) $\Delta(a = a)$ (axiom); (3) $\lambda x[\Delta(x = a)]a$ (from (2)); (4) $\lambda x[\Delta(x = a)]b$ (from (1) and (3) by Leibniz's Law); (5) $\Delta(a = b)$ (from (4) by λ-elimination).

[9] See also my "On Vague Objects, Fuzzy Logic, and Fractal Boundaries," *Southern Journal of Philosophy*, Supplementary Volume XXXIII (1994): 83-96.

[10] For example, Broome, Johnsen, Parsons.

To fill in some formal detail, let every statement under consideration have a value lying in the closed interval [0, 1] of the real numbers. ('Closed' means simply that the end-points 0 and 1 are included in the interval.) I shall call 0 and 1 the *integral* values. 0 represents complete or determinate falsehood, 1 complete or determinate truth. Identity statements are, of course, among those which may take non-integral values.

Truth conditions are given for compound statements in terms of arithmetical operations on values. For example:

$$v(\sim A) = 1 - v(A)$$

('–' represents ordinary subtraction)

$$v(A \& B) = \min\ (v(A), v(B))$$

(min (i, j) is the smaller of the two numbers i and j)

$$v(A \vee B) = \max\ (v(A), v(B))$$

(max (i, j) is the larger of i and j)

$$v(A \rightarrow B) = 1 \text{ if } v(B) \geq v(A);\ v(A \rightarrow B) = 1 - (v(A) - v(B)) \text{ if } v(B) < v(A)$$
$$v(A \leftrightarrow B) = \min\ (v(A \rightarrow B), v(B \rightarrow A))^{12}$$

---

[11] "On Three-valued Logic" (1920), p. 88, and "Philosophical Remarks on Many-valued Systems" (1930), p. 166; both in L. Borkowski, ed., *Jan Lukasiewicz: Selected Works* (Amsterdam: North Holland, 1970).

[12] Lukasiewicz, "A Numerical Interpretation of the Theory of Propositions" (1922), in Borkowski; L.A. Zadeh, "Fuzzy Logic and Approximate Reasoning," *Synthese*, xxx (1975): 407-28.

(Lukasiewicz himself took the nonintegral values of his semantics to be degrees of probability rather than degrees of truth (*ibid.*, p. 130).)

It is natural to think of ΔA as being true (determinately) just in case A has an integral value and as being false (determinately) otherwise:

$v(\Delta A) = 1$ if and only if either $v(A) = 1$ or $v(A) = 0$
$v(\Delta A) = 0$ if and only if $0 < v(A) < 1$

It is equally natural to think of ∇A as being true (determinately) just in case A has a nonintegral value and as being false (determinately) otherwise:

$v(\nabla A) = 1$ if and only if $0 < v(A) < 1$
$v(\nabla A) = 0$ if and only if either $v(A) = 1$ or $v(A) = 0$

Notice that the delta operators are duals, as Evans requires.

Fuzzy logic's detractors (for example, Michael Tye[13]) scoff at the "excess precision" afforded by the presence of continuum-many degrees of truth. What, they ask, has this bizarre sensitivity got to do with actual reasoning in the real world? This is largely to miss the point. It is only in the maximally general mathematical theory that infinitely many values are considered. Compare this "objection" to the general mathematical theory of computation: What does a theory that countenances infinitely many machines with tapes of unbounded length have to do with actual computation in the real world? In practical applications—for example, a logic of vagueness for implementation in an artificial-intelligence reasoning program, or a fuzzy expert system for controlling some industrial process—only a finite number of nonintegral values will be employed. The general theory is an *idealization*, abstracted from actual engineering practice (or, as it was in the first instance, from intended practice). Real engineers typically use the nonintegral values to *rank* propositions according to their relative degree of truth, rather than to ascribe absolute degrees of truth. As J. A. Goguen[14] says, "we should not expect particular numerical values...to be meaningful (except 0 and 1), but rather their *ordering*" (*ibid.*, p. 332). The spectrum of nonintegral values is thought of as roughly (and fuzzily) partitioned into regions that bear linguistic labels, such as 'very true', 'nearly true', 'more or less true', 'barely more true than false', 'more false than true', 'almost completely false', and so forth. Thus, to an engineer the assignments $v(p) = 0.9$, $v(q) = 0.95$ may indicate simply that p and q are both pretty true, and that q is a bit truer than p. Since the ranking is open, in the sense that more propositions may come along later and occupy intermediate positions, the engineer must have access to values *between* any two values, at any rate up to some resource limit. Generally, fuzzy logicians take the real line to be the appropriate idealization (although my own preference is for the rationals rather than the reals).

---

[13] "Vagueness: Welcome to the Quicksand," *Southern Journal of Philosophy*, Supplementary Volume XXXIII (1994): 1-22, here p. 11.

[14] "The Logic of Inexact Concepts," *Synthese*, XIX (1969): 325-73.

One reason for the suggestion that fuzzy logic is the logic of *vagueness*, rather than merely a logic of *inexactness*, is that in fuzzy logic the sorites paradoxes melt away.[15] The major premise of any sorites, while no doubt true enough for any conceivable practical purpose, is not completely true. Using the major premise once, or even several times, produces no difficulty. At each step, however, the statement that is inferred deviates from complete truth, and this deviation, while tiny at first, increases relentlessly with each application of modus ponens, until after enough iterations the conclusion is gapingly false.

This is illustrated by the classical paradox of the bald man (said to originate with Eubulides):

> (1) $\forall n$(x is not bald when he has n hairs on his head $\rightarrow$ x is not bald when he has n–1 hairs on his head)
> (2) x is not bald when he has a million hairs on his head
> $\therefore$ (3) x is not bald when he has 0 hairs on his head

(1) is not completely true (for the sorites argument reduces to absurdity the supposition that (1) is completely true).[16] Let its value be $1 - \delta$, for some $\delta > 0$. The idea is that for practical purposes, δ is negligible. Let us assume, for the moment, that the value of each elimination instance of a universal quantification, $\forall xP$, is identical to the value of $\forall xP$ itself. I shall represent 'x is bald when he has n hairs on his head' by 'Bn'. The value of the conclusion of the first step of the sorites is obtained by solving the equation:

---

[15] An example of a statement that is inexact but not vague is one recording a numerical measurement that is accurate only to two significant figures, the quantity measured being continuous. For a representation to be vague, there must be borderline cases of its applicability.

[16] As is well known, the remaining position, that (1) is completely false, has a high price. For then: $\exists n$(x is not bald when he has n hairs on his head and x is bald when he has n–1 hairs on his head) is completely true. The supposition that there is such an n strikes many as fantastic.

$$1 - \delta = 1 - (1 - v({\sim}B999{,}999))$$

and of the second step, ${\sim}B999{,}998$, by solving:

$$1 - \delta = 1 - ((1 - \delta) - v({\sim}B999{,}998))$$

and of the third step by solving:

$$1 - \delta = 1 - ((1 - 2\delta) - v({\sim}B999{,}997))$$

and so on. The values of the members of the sorites sequence ${\sim}B1{,}000{,}000$, ${\sim}B999{,}999$,... descend steadily toward 0, each decrement being by an amount that is locally negligible:

${\sim}B1{,}000{,}000$	1
${\sim}B999{,}999$	$1 - \delta$
${\sim}B999{,}998$	$1 - 2\delta$
${\sim}B999{,}997$	$1 - 3\delta$
.	.
.	.
.	.

Each iteration of modus ponens is truth preserving *enough* in that the value of the conclusion always differs by only a negligible amount from the value of the conjunction of the premises. But chain enough steps together, and a quantity that is negligible after a single iteration will start to make its presence felt. This strikes me as a revealing theoretical reconstruction of the sorites phenomenon. It fits well with what one is inclined to say pretheoretically about a sorites: each individual step looks all right, yet somehow disaster results.

At any rate, those are the basic ideas. The story is not yet completely correct, for it is surely counterintuitive that ${\sim}B999{,}999$ should be ranked as *less true* than ${\sim}B1{,}000{,}000$. These two statements are both completely true: neither is truer than the other, not even by a tiny amount. That the foregoing story holds otherwise is a consequence of the simplifying assumption that the value of *every* elimination instance of $\forall n({\sim}Bn \rightarrow {\sim}Bn{-}1)$ is $1 - \delta$. More plausibly, restricting the quantifier in different ways will produce formulas of differing values. In particular, there will be formulas of the form $\forall n{>}j({\sim}Bn \rightarrow {\sim}Bn{-}1)$ which are completely true. Of course, it is in the nature of things that we cannot specify the least j for which this is so, and nor can we specify the greatest k such that $v(\forall n{<}k({\sim}Bn \rightarrow {\sim}Bn{-}1)) < 1$. The graph of $v({\sim}Bn)$ against n consists of a number of segments of different gradients. Up in the high figures the line is flat (certainly the value of ${\sim}B1{,}000{,}000$, ${\sim}B999{,}999$, ${\sim}B999{,}998$ and ${\sim}B999{,}997$ is in each case 1). No segment quite meets its neighbor:

the point $(0, \sim B0)$ excepted, lines begin and end in a blur. Within these fuzzy regions there are no crisp facts of the matter concerning the values or the derivatives of the function $f(n) = v(\sim Bn)$ (unless one is an epistemic theorist, in which case the segments of the graph do meet, but always beneath a blot of ignorance). These fuzzy regions on the graph are regions of *higher-order* indeterminacy, a matter to which I shall return shortly.

The idea of using fuzzy logic to resolve sorites paradoxes has appealed to a number of recent writers. There seems to be a broad consensus among these writers that—contrary to what I am suggesting here—the precise culprit is modus ponens; the inference:

$$\sim Bn \rightarrow \sim Bn-1, \sim Bn \vdash \sim Bn-1$$

is invalid, it is said.[17] The thought is that an inference is valid if and only if the value of the conclusion can be *no less than* the value of the weakest premise; and, as we have seen, there are many points in the sorites sequence at which the value of the premise $\sim Bn$ exceeds the value of $\sim Bn-1$. One advocate of fuzzy logic who has objected to this treatment of modus ponens is Dorothy Edgington. She sensibly remarks: "To say that [modus ponens] is invalid obscures an important distinction: between the case where the fall in value of the conclusion is constrained by the values of the premisses, and the case where it is not" (*op. cit.*, p. 200).[18]

The all-or-nothing distinction between the logically perfect and the logically worthless that motivates the classical account of argument validity is completely foreign to fuzzy logic. Here, the engineer's principle of "right is right enough" holds sway. Each application of modus ponens in the sorites argument is good enough, since the extent to which the value of the conclusion deviates below the value of the weakest premise never exceeds δ, and by hypothesis δ is negligible. In general, where $A \rightarrow B$ is nearly true, B's value, although less than A's, will always be nearly the same as A's. Modus ponens is truth preserving enough.

---

[17] See, for example, G. Forbes, *The Metaphysics of Modality* (New York: Oxford, 1985), pp. 171-72, and *Modern Logic* (New York: Oxford, 1994), pp. 355-56; C. Peacocke, "Are Vague Predicates Incoherent?" *Synthese*, XLIV (1981): 121-41, p. 127; R.M. Sainsbury, *Paradoxes* (New York: Cambridge, 1988), pp. 42-43. Sainsbury there maintains that the fuzzy logician "must challenge the validity of modus ponens" (p. 42). In the second edition of the book (1995), he says (persuaded by D. Edgington, "Validity, Uncertainty, and Vagueness," *Analysis*, LII (1992): 193-203): "degree theory need give us no reason to say that modus ponens is not valid" (p. 42). He still thinks, however, that any fuzzy logician who does deny the validity of sorites reasoning (as opposed to denying the soundness of the reasoning on the ground that the premises are not fully acceptable) must deny the validity of modus ponens (pp. 41-42). On the analysis I present below, it is the invalidity of the classical rule of contraction, and not of modus ponens, that accounts for the invalidity of the sorites.

[18] See also her "Wright and Sainsbury on Higher-order Vagueness," *Analysis*, LIII (1993): 193-200.

In fuzzy logic, the unrestricted iteration of valid inferences may lead to trouble. A detailed syntactical analysis of the reasoning involved in the paradox of the bald man will both pinpoint the problem and afford a way of dealing with it. (The semantics of the situation must await another occasion.) I shall use the proof-theoretic apparatus introduced by Gerhard Gentzen,[19] and in particular his cut rule, or transitivity rule. A Roman letter represents a single formula and a Greek letter represents either a single formula, or a string of formulas separated by commas, or the null string.

$$\text{Cut} \qquad \frac{\Gamma \vdash \Theta, A \qquad\qquad A, \Pi \vdash \emptyset}{\Gamma, \Pi \vdash \Theta, \emptyset}$$

Premise (1) of the sorites argument, the major premise, will be abbreviated by the letter '*M*'. The first iteration in the chain of reasoning can be represented like this:

$$\begin{array}{rl} M \vdash {\sim}B1{,}000{,}000 \rightarrow {\sim}B999{,}999 & \\ {\sim}B1{,}000{,}000 \rightarrow {\sim}B999{,}999, {\sim}B1{,}000{,}000 \vdash {\sim}B999{,}999 & \\ M, {\sim}B1{,}000{,}000 \vdash {\sim}B999{,}999 & \text{by cut} \end{array}$$

The second iteration is:

$$\begin{array}{rl} M \vdash {\sim}B999{,}999 \rightarrow {\sim}B999{,}998 & \\ {\sim}B999{,}999 \rightarrow {\sim}B999{,}998, {\sim}B999{,}999 \vdash {\sim}B999{,}998 & \\ M, {\sim}B999{,}999 \vdash {\sim}B999{,}998 & \text{by cut} \end{array}$$

Amalgamating the results of these two iterations by a further application of cut produces:

$$M, M, {\sim}B1{,}000{,}000 \vdash {\sim}B999{,}998[20]$$

Notice that the cut rule is very picky. It insists that Γ and Π, whatever they may be, are written out in full on the left hand side of the line that is being derived. The rule makes no provision for doing away with the second occurence of *M* so generated. In classical logic, any repetitions of a formula, once obtained, may be deleted by an application of the separate rule of *contraction*:

---

[19] "Investigations into Logical Deduction" (1934), in M.E. Szabo, ed., *The Collected Papers of Gerhard Gentzen* (Amsterdam: North-Holland, 1969).

[20] Aficionados of the sequent calculus will see that exchanges are being suppressed (for simplicity of presentation).

$$\frac{A, A, \Gamma \vdash \Delta}{A, \Gamma \vdash \Delta}$$

This cannot be done here, however. The fuzzy logic that I am advocating is a *contraction-free* logic.

At the end of the chain of inferences that make up the sorites, we arrive at:

$$M, M, \ldots, M, {\sim}B1{,}000{,}000 \vdash {\sim}B0$$

(To be precise, there are 1,000,000 occurrences of the letter '*M*'.) In the absence of unrestricted contraction, this cannot be transformed into the unwanted:

$$M, {\sim}B1{,}000{,}000 \vdash {\sim}B0$$

Think of it this way. The number of occurences of '*M*' is a record of the number of iterations of modus ponens. Where the number of iterations is small, contraction may safely be applied (always assuming the situation is such that each iteration is roughly truth preserving, in the sense explained previously: the value of the derived formula never deviates by more than some small fixed amount from the value of the conjunction of the two formulas from which it is derived). But where a statement of derivability is flagged with a long sequence of repetitions, contraction is dangerous.

The flagging formula, in this case *M*, need not be a universal quantification. For example, a conjunction of conditionals will serve just as well. The point is that applications of cut can be used to chain together a series of instances of modus ponens only if some formula is found with which progressively to flag each successive result of applying cut.[21] One notationally economical way to achieve the flagging in the case where the sorites reasoning has a series of conditional premises in place of *M* is to use the antecedent of the first application of modus ponens as the flagging formula. (The conditional itself would serve instead, but is longer.) The idea is that *P* is used once in the inference:

$$P, P \rightarrow Q \vdash Q$$

and then *again* in using *Q* in the inference:

---

[21] Gentzen's rule of *weakening* may be used to introduce occurrences of the flagging formula: $\frac{A, \Gamma \vdash \Delta}{A, A, \Gamma \vdash \Delta}$

$Q, Q \rightarrow R \vdash R$

and so on. Thus combining these two inferences in a way that keeps track of the number of times P is used will yield:

$P, P, P \rightarrow Q, Q \rightarrow R \vdash R$

(Imagine that whenever a formula is employed as a premise of a step in the reasoning, it itself, together with each formula higher in the derivation upon which it depends, lights up. The notation keeps track of the number of times that P lights up.) In a contraction-free logic, these repetitions of P, once introduced, can be eliminated only in special circumstances.[32]

Contraction-free logics were first studied by Arthur Prior and Carew Meredith in the 1950s. A punting trip in Oxford in the summer of 1956 was the start of a collaboration between Prior, Meredith, his cousin David Meredith, John Lemmon, and Ivo Thomas, which ensued in the legendary joint paper, "Calculi of Pure Strict Implication." This circulated in dittoed form for many years.[33] In 1963, a second paper appeared, "Notes on the Axiomatics of the Propositional Calculus," written by Prior and Carew Meredith.[34] Section 7 of the paper gives details of a contraction-free combinatory logic first proposed by Meredith in 1956 or earlier.[35] In their romp through weak implicational logic, Prior and Meredith doubtless gave little thought to practical applications of the systems they discovered.[36] Nowadays, contraction-free logic is the object of considerable attention in computer science, largely because of the remarkable fact that, if contraction is forbidden, then first-order predicate logic is decidable.[27] Among the computer-science community, probably the best-known contraction-free logic is J. Y. Girard's[28] linear logic. Girard aptly describes the contraction rule as permitting the use of resources ad libitum: in the presence of contraction, a premise P can be used any number of times in the course of obtaining a derivability statement that shows only *one* occurrence of P on the left.[29] Various other contraction-free logics are under study, particularly in Australasia, notably by Ross Brady, by Prior's former student Robert Bull, and by Martin Bunder.[30] Fuzzy logic is a pleasant addition to the family.

---

[32] A crude first approximation to the semantics of the situation is this. The value of the antecedent of any sequent is the minimum of the values of the individual antecedent formulas *except that* before the minimum is calculated the value of any repeated formula is multiplied by itself n times, where n is the number of occurrences among the antecedent formulas of the formula in question. For example, the value of the antecedent of the sequent $p, p, q \vdash r$ is min $(v(p) \times v(p), v(q))$.

[33] It was eventually published in J.W. Davis, D.J. Hockney, W.K. Wilson, eds., *Philosophical Logic* (Dordrecht: Reidel, 1969), pp. 215-50.

[34] *Notre Dame Journal of Formal Logic*, IV (1963): 171-87.

[35] An unpublished typescript of Prior's gives the date of Meredith's system as '≤1956'.

[36] It is interesting that, in addition to the various applications of contraction-free logic which have subsequently arisen, the condensed detachment rule that Meredith devised (Meredith and Prior, *op. cit.*) is the basis of the Resolution Principle employed in most current approaches to automated theorem proving—M. Bunder, "Logics without Contraction II"; see also my "Prior's Life and Legacy"; both are in my *Logic and Reality: Essays on the Legacy of Arthur Prior* (New York: Oxford, 1996), pp. 337-49 and 1-40.

Turning once more to the issue of vague identity and Evans's derivation, the first point to notice is that in fuzzy logic, inference by Leibniz's Law is not always correct. Where $a = b$ is indeterminate, an inference:

$$\emptyset(a) \vdash \emptyset(b)$$

may fail, for the value of $\emptyset(a)$ may be 1 and yet the value of $\emptyset(b)$ be nonnegligibly less than 1. Evans's own references to Leibniz's Law seem to have been to the conditionalized form:

$$a=b\rightarrow(\emptyset(a)\leftrightarrow\emptyset(b))$$

for he speaks, in a passage quoted earlier, of strengthening Leibniz's Law with a 'Definitely' prefix.[31] But there is nothing law-like about this conditional in the present semantics. When the value of '$a = b$' drops below 1 the value of:

$$\lambda x[\nabla\ (x = a)]a\leftrightarrow\lambda x[\nabla\ (x = a)]b$$

---

[27] See, for example, J. Ketonen, R. Weyhrauch, "A Decidable Fragment of Predicate Calculus," *Theoretical Computer Science*, xxxii (1984): 297-307. The reason is that Gentzen's subformula property yields a decision method for predicate logic provided there is a known bound to the length of the sequents involved in a cut-free proof of any given formula, and in the absence of contraction this bound is simply the length of the formula to be proved. (The subformula property is as follows: the only formulas that occur in a cut-free proof of any given sequent $\Gamma \vdash \Pi$ are subformulas of the formulas in Γ and Π (Gentzen, pp. 87-88).) Incidentally, there is no presumption that the cut rule can be *eliminated* from the proof theory of the fuzzy logic here considered.

[28] "Linear Logic," *Theoretical Computer Science*, L (1987): 1-102.

[29] Girard, "Geometry of Interaction" (unpublished, 1988), p. 13.

[30] Brady, "Gentzenization and Decidability of Some Contraction-less Relevant Logics," *Journal of Philosophical Logic*, xx (1991): 97-117; R. Bull, "Logics without Contraction I," in my *Logic and Reality*, pp. 317-36; Bunder, *op. cit.*

[31] Evans's intention, in the passage quoted, seems to have been to distribute Δ across the conditional. If so, he was in error, for

$$\Delta(a = b\rightarrow(\emptyset(a)\leftrightarrow\emptyset(b)))\rightarrow(\Delta(a = b)\rightarrow\Delta(\emptyset(a)\leftrightarrow\emptyset(b)))$$

can take the value 0.

becomes 0 (since the value of the left-hand formula is 0 and the right hand formula 1), and so the value of:

$$a = b \rightarrow (\lambda x[\nabla(x = a)]a \leftrightarrow \lambda x[\nabla(x = a)]b)$$

becomes nonintegral. If the value of '$a = b$' rides arbitrarily close to 1, the value of this instance of the "law" rides arbitrarily close to 0.

This, then, is the fallacy in Evans's derivation. Although lines (2) and (4) are, we are supposing, completely true, the inference to (5):

$$\lambda x[\nabla(x = a)]b, \sim\lambda x[\nabla(x = a)]a, a = b \rightarrow (\lambda x[\nabla(x = a)]a \leftrightarrow \lambda x[\nabla(x = a)]b) \vdash a \neq b$$

can deliver a conclusion whose value is arbitrarily close to zero.

It would be nice if things could be left exactly there. Unfortunately, this treatment of Evans's derivation is open to a powerful—although, in the end, ineffective—objection. Forms of the objection have been developed by B. Garrett and H. W. Noonan.[*] (Neither Garrett nor Noonan was writing concerning fuzzy logic.) The objection is as follows.

The foregoing criticism of Evans's derivation concedes that lines (2) and (4) of the derivation are true. Moreover, these formulas are not merely true to a high degree, but determinately true, for they are derived from the assumptions '$\nabla(a = b)$' and '$\sim\nabla(a = a)$', respectively, and these two formulas in turn, if true, are determinately true, since in the semantics all formulas of the form '∇A' are integral valued. In other words, given only the innocuous assumption that the operation of λ-abstraction preserves the value of the formula to which it is applied, the assumption that $\nabla(a = b)$ entails that it is determinately true that b has the property $\lambda x[\nabla(x = a)]$, and the assumption that $\sim\nabla(a = a)$ entails that it is determinately true that a lacks that same property. So does it not further follow, just as Evans thought, that '$a = b$' is determinately false? For surely the proponent of fuzzy identity means to count '$a = b$' as indeterminate just in the case where it is indeterminate whether a and b have all their properties in common? If it is determinately true that a and b do *not* have all their properties in common, then how can even a fuzzy logician draw back from the conclusion that '$a = b$' is determinately false?

In short, while it is true that what Evans calls Leibniz's Law is of no help in moving from (2) and (4) to (5), there is (the objection continues) a related principle that will take us directly from (2) and (4) not only to (5) but also to Evans's (5′), $\Delta(a \neq b)$, namely, the principle that if, determinately, *b* has some property that, determinately, *a* lacks, then, determinately, *a* and *b* are distinct. In symbols:

---

[*] Garrett, "Vague Identity and Vague Objects," pp. 347-48; Noonan, *Personal Identity* (New York: Routledge, 1989), section 6.6, and "Vague Identity Yet Again," *Analysis*, L (1990): 157-62, pp. 160-62.

$$\sim\emptyset(a), \emptyset(b), \Delta(\sim\emptyset(a)), \Delta(\emptyset(b)) \vdash a \neq b \,\&\, \Delta(a \neq b)$$

(Noonan calls a weaker form of this "the principle of the Diversity of the Definitely Dissimilar."[33] The present principle is that of the Definite Diversity of the Definitely Dissimilar.) This objection fails, as I shall explain.[34]

A vague predicate such as 'bald' effects a tripartite division of the objects in its range of significant application, dividing them into the positive extension, those things which are definitely bald, the negative extension, those things which are definitely not bald, and the borderline cases. None of these divisions is sharp. There is no more a sharp line between the positive extension of 'bald' and the border region than there is between the positive extension and the negative extension. Bertrand Russell[35] remarked on this in 1923 (and it has been emphasized recently by several writers):

> [W]ords are attributable without doubt over a certain area, but become questionable within a penumbra, outside which they are again certainly not attributable.... [T]he penumbra itself is not accurately definable, and all the vaguenesses which apply to the primary use of words apply also when we try to fix a limit to their indubitable applicability (*ibid.*, p. 87).

As it is sometimes put, there are borderline cases of the predicate 'is a borderline case of the predicate *bald*'.[36] This higher-order indeterminacy or higher-order vagueness has already been mentioned in the case of the graph of $v(\sim Bn)$ against n: the boundaries between the integral-valued sentences and the indeterminate sentences are themselves indeterminate.

---

[33] "Vague Identity Yet Again," p. 160.

[34] In a response to Noonan, Keefe—"Contingent Identity and Vague Identity," *Analysis*, LV (1995): 183-90—objects to Evans's derivation on the basis of a claim that delta-predicates do not denote properties. Her grounds for holding the latter are seemingly ad hoc:

> If we allow that it can be genuinely indeterminate whether something has a particular property, then we must deny that 'it is indeterminate whether...' denotes a further property. Assuming that it does, or equivalently assuming that it can be substituted into the Diversity of the Dissimilar, begs the question against this possibility (p. 188).

If the present treatment is correct, then one can allow both that it is indeterminate whether something has a particular property *P* and that $\lambda x[\nabla Px]$ is genuinely a property.

[35] "Vagueness," *Australasian Journal of Psychology and Philosophy*, I (1923): 84-92.

[36] The matter is discussed by, for example, Dummett in "Wang's Paradox"; D. Hyde, "Why Higher Order Vagueness Is a Pseudo-Problem," *Mind*, CIII (1994): 35-41; Sainsbury, "Concepts without Boundaries," Inaugural Lecture, 1991, King's College, London, and "Is There Higher-Order Vagueness?" *Philosophical Quarterly*, XLI (1991): 167-82; T. Williamson, *Vagueness* (New York: Routledge, 1994); C. Wright, "Further Reflections on the Sorites Paradox," *Philosophical Topics*, XV (1987): 227-90, and "Is Higher-order Vagueness Coherent?" *Analysis*, LII (1992): 129-39.

A semantics for vagueness that fails to acknowledge the indeterminacy of the penumbra is at best incomplete. As Sainsbury puts it, "no sharp cut-off to the shadow of vagueness is marked in our linguistic practice, so to attribute it to [a vague] predicate is to misdescribe it."[37] Sainsbury believes that this presents a fundamental difficulty for fuzzy logic:

> [T]his...is what scuppers the...descriptions of vague languages offered by fuzzy logicians.... [T]he fuzzy logician, too, whether he likes it or not, will be committed to a threefold partition: the sentences which are true to degree 1, those true to degree 0, and the remainder. But to what in our actual use of language does this division correspond? It looks as if, as before, it should correspond to the sentences true beyond the shadow of vagueness, those in some kind of borderline position, and those false beyond the shadow. But...we do not know, cannot know, and do not need to know these supposed boundaries to use language correctly. Hence they cannot be included in a correct description of our language (*ibid.*, pp. 11-12).

Kit Fine,[38] too, objects that fuzzy logic cannot "express the fact that Herbert [is] a *clear* borderline case of a bald man" (*ibid.*, p. 293). Yet I think fuzzy logicians will see little force in such criticisms. As the example of the graph of $v(\sim Bn)$ against n may make clear, fuzzy logic is certainly not committed to the idea that there is a sharp boundary between the sentences of value 1 and the sentences of value less than 1.

What is desirable is some way of mathematicizing the—as it stands—highly *un*mathematical notion of a graph with fuzzy spots. One way of proceeding is to attempt to formalize the idea, previously mentioned, that the spectrum of nonintegral values is to be thought of as fuzzily partitioned into regions labeled 'nearly true', 'almost completely false', and so on. Let us take these routine English phrases seriously, and let us call them and their cognates *fuzzy truth values*. As I have presented things thus far, the fuzzy truth values lie outside the formal semantics: in the mathematical theory, we find only the integral and the nonintegral values, in themselves perfectly sharp. The fuzzy values enter only at the level of informal talk about the mathematical theory, as in "statements with values around 0.8 or better are approximately true and those with values around 0.95 or better are almost completely true." Can the fuzzy truth values be brought within the ambit of the formal theory itself?

---

[37] "Concepts without Boundaries," p. 11.

[38] "Vagueness, Truth, and Logic," *Synthese*, xxx (1975): 265-300.

L. A. Zadeh[39] has attempted to do exactly that. His proposal—and perhaps the best way to view it is as being in the spirit of the set-theoretic reduction of numbers to sets—is to allow fuzzy sets to do duty for fuzzy truth values. A fuzzy set is a set of pairs $<x, n>$, where x is an object and $0<n\leq 1$; n represents x's "degree of membership" in the fuzzy set.[40] Each fuzzy truth value is identified with a fuzzy subset of the set consisting of the integral and nonintegral values (which is to say, is identified with a fuzzy subset of the interval [0, 1]). For example, suppose the fuzzy truth value 'approximately true' is identified with a fuzzy set α. Then the degree of membership in α of some nonintegral value i signifies the extent to which being true to degree i counts as an instance of being approximately true. We may expect that $<0.95, 1>$ is in α and that $\forall i(<0, i>\notin \alpha)$.

This set-theoretic reduction of the fuzzy truth values is not entirely satisfying. The resulting semantics hovers between pure and applied.[41] The problem of formalizing the fuzzy truth values in a philosophically satisfying way is the principal outstanding problem of fuzzy logic, and is a challenge to philosophical logicians working on the theory of vagueness. Nevertheless, Zadeh's set-theoretic reduction of the fuzzy values amply illustrates the point that fuzzy semantics is far from committed to there being a sharp boundary between complete truth and less-than-complete truth; for two fuzzy sets can shade into one another. Points in the border region belong to both sets, though generally to different degrees. In particular, the fuzzy set 'completely true' shades into the fuzzy set 'almost completely true'. The values 1 and 0.95, for instance, will belong to both sets, though to different degrees. Fuzzy logic need impose no sharp cut-off upon the "shadow of vagueness."

I shall say no more here concerning Zadeh's project of mathematicizing the fuzzy truth values. The immediate problem is the semantics of the delta operators. A statement to the effect that such-and-such a statement is indeterminate may itself be indeterminate, and likewise for a statement to the effect that such-and-such a statement is determinate. Yet in the foregoing semantics, statements of the form ∇A and ΔA always take either the value 0 or the value 1, which is to say, are determinate. If the semantical clauses for Δ and ∇ are to save the phenomenon of higher-order indeterminacy, they must allow statements of these forms to sometimes take nonintegral values. Then the distinction between a's being a clear borderline case of a predicate F and a's being a borderline borderline case of F can be expressed in terms of the distinction between $v(\nabla Fa)$ being 1 and $v(\nabla Fa)$ being nonintegral. Here is one way of achieving this.

---

[39] "Fuzzy Logic and Approximate Reasoning," *Synthese*, XXX (1975): 407-28; R.E. Bellman and Zadeh, "Local and Fuzzy Logics," in J.M. Dunn, G. Epstein, eds., *Modern Uses of Multiple-Valued Logic* (Dordrecht: Reidel, 1977).

[40] Zadeh, "Fuzzy Sets," *Information and Control*, VIII (1965): 338-53.

[41] See my "On When a Semantics Is Not a Semantics," *Journal of Philosophical Logic*, VIII (1979): 399-413, "Pure Semantics and Applied Semantics," *Topoi*, II (1983): 197-204, "What Is a Semantics for Classical Negation?" *Mind*, XCV (1986): 478-90, "Vagueness and Bivalence," *Proceedings of the Aristotelian Society*, LXVIII (1994): 193-200; and Lewis, "Relevant Implication," *Theoria*, LIV (1988): 161-74, p. 173.

With each statement A is associated A's *higher-order profile*, written $\Omega(A)$. $\Omega(A)$ is a subset of $[0, 1]\times[0, 1]$, which is to say, is a set of ordered pairs, each member of each pair being either 1, 0, or nonintegral. The idea is that each member $<n, i>$ of $\Omega(A)$ records the degree i to which it is true that A possesses the value n. I shall call the right-hand members of pairs in a profile *higher-order degrees* and the left-hand members *primary degrees*. (All the remarks made earlier concerning "excess precision," idealization, ranking, and so forth apply equally to these new higher-order degrees.) Where A definitely has the value n, then $<n, 1> \in \Omega(A)$. $\Omega(A)$ respects the conditions:

1. $\forall i \forall j \forall k(<i, j> \in \Omega(A) \rightarrow (<i, k> \in \Omega(A) \rightarrow j = k))$
2. $\forall i \forall j \forall k(<i, 1> \in \Omega(A) \rightarrow (<j, k> \in \Omega(A) \rightarrow <j, k> = <i, 1>))$

If there is an n such that $<n, 1> \in \Omega(A)$, $\Omega(A)$ is said to be *unitary*. Where it is indeterminate whether A has a certain value n, this is marked by the presence in the higher-order profile of a pair $<n, i>$, where i is nonintegral. The profile may contain a number of such pairs, each containing a different primary value.

The "fuzzy spots" or regions of higher-order indeterminacy on the graph of $v(\sim Bn)$ against n are regions where $\Omega(\sim Bn)$ is not unitary. For example, let us pick numbers x, y, and z $(x<y<z)$ such that x and y lie in a fuzzy region and z lies clearly outside it on a crisp line. $\Omega(\sim Bx)$ is, say, $\{<1, 0.8>, <0.9, 0.9>\}$, $\Omega(\sim By)$ is $\{<1, 0.95>\}$, and $\Omega(\sim Bz)$ is $\{<1, 1>\}$. At $n=z$ the line is definite. At $n=y$ it is not so definite, drawn fainter perhaps, while at $n=x$ the line has disappeared altogether.

There are three cases to be considered in reformulating the semantical clauses for Δ and ∇.

Case 1. $\Omega(A)$ is unitary. The member of $\Omega(A)$ is either $<1, 1>$ or $<0, 1>$ or $<i, 1>$ for some nonintegral i. In other words, A is either determinately determinate or determinately indeterminate. As before, where the primary degree is integral, $v(\Delta A)=1$ and $v(\nabla A)=0$, and where the primary degree is nonintegral, $v(\nabla A)=1$ and $v(\Delta A)=0$.

Case 2. $\Omega(A)$ is nonunitary and none of the primary degrees in $\Omega(A)$ is integral. Here it is indeterminate which primary degree A has, but it is not indeterminate whether or not A is indeterminate. $v(\nabla A)=1$ and $v(\Delta A)=0$.

Case 3. $\Omega(A)$ is nonunitary and at least one of the primary degrees in $\Omega(A)$ is integral. Here it is indeterminate whether A is determinate. Either $<1, i> \in \Omega(A)$ or $<0, j> \in \Omega(A)$ or both are, for some nonintegral i and j. If only the first of these pairs is present then $v(\Delta A)=i$, and if only the second is present $v(\Delta A)=j$. If both are present, $v(\Delta A)=\max(i, j)$. $v(\nabla A)$ is $1 - v(\Delta A)$ (thus, for example, if it is almost completely true that A is determinate, then it is almost completely false that A is indeterminate). Assuming that $v(\Delta \sim A)=v(\Delta A)$, this treatment of ∇ preserves duality.

Each $\Omega(A)$ considered so far consists of ordered pairs. Higher-order profiles of this sort are called *two-dimensional*. As promised, the use of two-dimensional higher-order profiles fuzzifies the division between the determinate and the indeterminate. Still higher-order divisions remain unfuzzy. For example, the two-dimensional apparatus leaves sharp the division between those sentences which are determinately indeterminate and those which are indeterminately indeterminate. This division, too, can be fuzzified if desired, at the price of adding one further dimension to each higher-order profile. A three-dimensional higher-order profile consists of triples $<n, i, j>$. j is the degree to which it is true that i is the degree to which it is true that the sentence in question possesses the value n. This process can be continued indefinitely. Adding a fourth dimension fuzzifies third-order divisions, such as that between the determinately indeterminately indeterminate and the indeterminately indeterminately indeterminate. Where no such boundary is to remain sharp the members of the higher-order profiles are infinite vectors $<x_1, x_2 ... x_n, ...>$ (and orders of indeterminateness that are even higher still can be coded into a profile, if desired, by employing vectors of increasing order type). It is, of course, only the finite cases that are of use to engineers, and how low a dimensionality one can get away with will depend on the software project in hand. There is no point in paying for dimensions that serve only to fuzzify boundaries too fine to be of any practical relevance.

Let us return to the discussion of Evans's argument and recall that the crucial step in the objection from the principle of the Definite Diversity of the Definitely Dissimilar, above, was the claim that the proof gives us a property, $\lambda x[\nabla\ (x=a)]$, which, determinately, b has and a does not have. As Noonan puts it:

> What the Evans...argument demonstrates is that indefinite identity, so conceived, can have no application; in the case in which '$a=b$' is indeterminate in truth-value...there will be a predicate 'it is indeterminate whether $x = a$' definitely true of b, whose negation is definitely true of a.[42]

Noonan accuses critics of Evans's argument of rejecting the principle that if "there is a predicate definitely true of b whose negation is definitely true of a [then it is not] indeterminate whether a is b," saying that the rejection of this principle is "the *only* road forward for an opponent of Evans's position."[43]

The fallacy is plain. Once higher-order indeterminacy is acknowledged, one cannot rely on an indeterminate statement's being determinately indeterminate. (So here is a second reason why Evans was mistaken in his claim that Δ and ∇ "generate a modal logic as strong as S5": ∇A does not imply $\Delta\nabla A$.[44]) That it is indeterminate whether $a=b$ does *not* entail that there is "a predicate...definitely true of b, whose negation is definitely true of a." The statement '$\nabla(a=b)$' does entail that $\lambda x[\nabla\ (x=a)]b$, but both these statements may themselves be indeterminate. Critics of Evans's argument certainly need not reject the principle that Noonan states. In my view, the principle of the Definite Diversity of the Definitely Dissimilar is self-evidently true (whatever that means), but it is of no more help than Leibniz's Law in the attempt to justify the move from (2) and (4) to (5) and (5′).

If Evans's derivation is enriched with the further assumption that '$\nabla(a=b)$' is determinate, and the principle of the Definite Diversity of the Definitely Dissimilar is employed, then the derivation of (5) and (5′) may seem to go through successfully.[45] Discharging the new assumption by reductio ad absurdum produces a derivation of:

---

[42] *Personal Identity*, pp. 135-36. For reasons of uniformity, I have replaced 'it is indeterminate whether $x = b$', which occurs in the original, by 'it is indeterminate whether $x = a$', and I have replaced the immediately following occurence of 'a' by 'b' and the immediately following occurence of 'b' by 'a'. I have made similar replacements in the next quotation.

[43] "Vague Identity Yet Again," pp. 160-62.

[44] See also Garrett, "Vagueness, Identity, and the World," p. 354. As Dummett pointed out as early as 1975, the logic will be weaker than S4, since $\Delta\ A$ does not imply $\Delta\ \Delta\ A$ ("Wang's Paradox," p. 311). As I have already mentioned, neither the T-axiom $\Box A\rightarrow A$ nor the K-axiom $\Box(A\rightarrow B)\rightarrow(\Box A\rightarrow\Box B)$ holds when '$\Box$' is replaced by 'Δ'.

[45] For some doubts of a different nature, see my "On Vague Objects, Fuzzy Logic, and Fractal Boundaries."

$$\nabla(a=b) \vdash {\sim}\Delta\, \nabla(a=b)$$

If the derivation holds good, this is an interesting and important result (and perhaps Evans would have been pleased enough to emerge as discoverer of a new property of vague identity[46]). Employing the usual metaphors, the result tells us that, if a predicate '$x = a$' is vague, then the 'shadow of vagueness' that it casts is *all* penumbra: all the borderline cases are borderline borderline cases.[47]

B. JACK COPELAND

University of Canterbury/New Zealand
and University of Sydney/Australia

---

[46] See Burgess, "Vague Identity: Evans Misrepresented."

[47] The strain on the traditional metaphor of a shadow is evident. Perhaps a better metaphor is provided by fractal surfaces such as the Sierpinski carpet—B.B. Mandelbrot, *The Fractal Geometry of Nature* (New York: Freeman, 1977, rev. and exp. 1982, 1983). This is a two-dimensional surface which is "all edge": like a shadow, the Sierpinski carpet covers a finite area, but every point on the carpet lies on an edge of the carpet. An interesting route of enquiry is to fuzzify the Sierpinski carpet.

THE LOT OF THE CAUSAL THEORY OF MENTAL CONTENT

The thesis of this paper is that the *causal theory of mental content* (hereafter CT) is incompatible with an elementary fact of perceptual psychology, namely, that the detection of distal properties generally requires the mediation of a "theory." I shall call this fact the *nontransducibility of distal properties* (hereafter NTDP). The argument proceeds in two stages. The burden of stage one is that, taken together, CT and the *language of thought hypothesis* (hereafter LOT) are incompatible with NTDP. The burden of stage two is that acceptance of CT requires acceptance of LOT as well. It follows that CT is incompatible with NTDP. I organize things in this way in part because it makes the argument easier to understand, and in part because the stage-two thesis—that CT entails LOT—has some independent interest and is therefore worth separating from the rest of the argument.

I. STAGE ONE: THE CONJUNCTION OF CT AND LOT IS INCOMPATIBLE WITH THE NONTRANSDUCIBILITY OF DISTAL PROPERTIES

Let us begin by clarifying some terms. By LOT, I mean the hypothesis that the human scheme of mental representation satisfies the following conditions:

(1) It has a finite number of semantically primitive expressions individuated syntactically.
(2) Every expression is a concatenation of the primitives.
(3) The content of any complex expression is a function of the contents of the primitives and the syntax of the whole expression.

The *locus classicus* of this thesis is Jerry Fodor's *The Language of Thought*.[1] I shall also use 'LOT' to refer to the language of thought itself rather than the hypothesis. The context will resolve any ambiguity this might otherwise introduce.

By CT, I mean the doctrine that the contents of the semantic primitives in the human scheme of mental representation are determined by their role in detection. The basic idea is just that the content of a primitive r in a system Σ is the property P if there is the right kind of causal connection between instantiations of P and tokenings of r by Σ's detectors. Causal theories come in a variety of fla-

---

[1] New York: Crowell, 1975.

0022-362X/97/9410/535–42

vors.[2] I am simply going to assume familiarity with these theories in what follows.[3] The only feature of the theory that will play a role in the discussion that follows is the following uncontroversial feature: assuming LOT, CT requires that every property represented by a primitive in LOT be detectable.

Now for NTDP. Distal properties generally cannot be directly transduced.[4] Instead, the detection of distal properties must be mediated by what we might as well call a theory about that property.[5] To detect cats (an instantiation of catness) requires a theory that says, in effect, what sorts of proximal stimuli are reliable indicators of catness. To detect cats visually, you have to know how cats look. The same goes for colors, shapes, and sizes: for these to be reliably detected visually under changes in perspective, lighting, and distance requires knowledge of such facts as that retinal-image size varies as the inverse square of the distance to the object.

Much of the knowledge[6] that mediates the detection of distal properties must be acquired: we are, perhaps, born with a tacit knowledge of Emmert's Law, but we are not born knowing how cats look, or with the ability to distinguish edges from shadows. We must, then, learn the theory that mediates cat recognition. Learning the theory will require formulating and confirming hypotheses such as these:

---

[2] See Dennis Stampe, "Towards a Causal Theory of Linguistic Representation," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy* (Minneapolis: Minnesota UP, 1977), pp. 42-63; Fred Dretske, *Knowledge and the Flow of Information* (Cambridge: MIT, 1981); Fodor, *A Theory of Content and Other Essays* (Cambridge: MIT, 1990); Ruth Millikan, *Language, Thought, and Other Biological Categories* (Cambridge: MIT, 1984); David Papineau, *Reality and Representation* (Cambridge: Blackwell, 1987).

Some readers might wonder at the inclusion of adaptationist theories like those of Papineau and Millikan as flavors of CT. These theories differ from other causal theories only in using adaptational role to select those causal connections that are content fixing.

[3] A review can be found in my *Meaning and Mental Representation* (Cambridge: MIT, 1989).

[4] "Transduced" in the sense of Z. Pylyshyn, *Computation and Cognition* (Cambridge: MIT, 1984), chapter 6. The central idea is that transduction is a cognitively primitive operation that maps physical events at the sensory surface onto computational events "upstream." It is controversial whether there *are* any transducers in this sense in human functional architecture. It is not controversial that transducers thus conceived are not responsible for distal property detection.

This "engineering" conception of transduction is to be distinguished from that found in the mathematical theory of computation where a transducer is just any finite automaton.

[5] It is controversial in psychology just what the form and exact content the mediating theory might turn out to have. It might turn out to be a point in weight space—for example, Paul M. Churchland, *A Neurocomputational Perspective* (Cambridge: MIT, 1989), chapters 9-11. According to LOT, if it is learned, it will be a set of sentences in LOT. If it is innate, it might, in some sense, be implicit in the architecture. For a review of some ways such information might be implicit, see my "Inexplicit Information," in M. Brand and R.M. Harnish, eds., *The Representation of Knowledge and Belief* (Tucson: Arizona UP, 1986), pp. 116-26.

[6] I use 'knowledge' here as it is used in cognitive psychology, which does not require that what is known be true or justified. Knowledge in this sense is information that is treated as if it were reliable by the system that uses it. This is what allows us to say that the visual system must know such things as that retinal-image size varies as the inverse square of the distance to the object even though, in the ordinary sense of 'know', this was not widely known in Europe until the seventeenth century.

(A) Cats have whiskers.
(B) Cats have four legs.
(C) Cats have fur.

According to LOT, these hypotheses are represented as sentences in LOT. As such, they require for their formulation, a symbol for cats, that is, a |cat|.[7] But, according to CT, you cannot have a |cat| until you have the ability to detect cats. According to psychology, you cannot have the ability to detect cats until you have a theory of cats. According to LOT, you cannot have a theory of cats until you have |cat|s. So, you cannot make the conjunction of LOT and CT compatible with psychology.

Objections and replies. Objection one. That is all right for |cat|s, but what about |square|s? What reason is there to believe that we have to learn a theory of squares to recognize them?

Reply one. Suppose the objection is right about squares and circles and some other stuff. What are you going to do about cats? Your only option is to suppose that |cat| is not a primitive symbol in LOT. But if it is not a primitive symbol in LOT, then it must be a complex symbol in LOT, that is, a Tarskian combination of other symbols in LOT. This implies that |cat| can be defined in LOT in terms of symbols for properties whose detection does not require a learned theory. Good luck: this kind of reductionism has a dismal history; if you want to revive it, you are on your own.

Reply two. The only argument on offer for the idea that one *does not* have to learn a theory of squares to recognize squares is that you *could not* learn it. Indeed, Fodor argues quite correctly in *The Language of Thought* that you could not learn to recognize any property represented by a primitive of LOT, the argument being, in effect, just the one given above. But this cuts as heavily against LOT as it does against the idea that squares can be detected without the aid of acquired information.

Objection two. The thing is done by degrees. You start out with some relation between distal stimuli and *R*. Since you do not yet know much about cats, *R* cannot mean catness. It means, as it might be, *p*catness ('*p*' for proximal). Since your theories have to use the symbols you have got, you have hypotheses like these: *p*cats have whiskers;[8] *p*cats have four legs; *p*cats have fur. As the theory gets elaborated, eventually, we are able to recognize cats reliably.

---

[7] A |cat| is a mental representation whose content is the property of being a cat.

Reply one. Why should the theory become elaborated? There is not, after all, an error signal that says, in effect, that the theory is not right yet. By hypothesis, the theory is, at any stage you like, as right as it could possibly be for *p*catness. This is because CT implies that you cannot have |*p*cat|s in the theory unless you have a good enough theory of *p*cats to underwrite the connection between *p*cats and some symbol in LOT. To elaborate the theory in the right direction, you have to know that the theory is not yet a good enough theory of catness, and that, it would seem, requires having |cat|s in your repertoire, applying them to various things—for example, *p*cats—and finding out that they are not cats.

It is easy to get confused about this, because one imagines someone who has not yet learned to recognize cats reliably looking at a cat and having the same percept as you and I would have. As George Berkeley would have said, he hath his eyes and the use of them as well as you and I; and surely that percept is a percept of a cat.

Quite right. But (a) it is not a |cat| since, as Berkeley also pointed out, it represents some particular cat of definite color, shape, and so on, whereas a |cat| represents any cat whatever; and (b) the structural similarities between the percept and the cat that make it so tempting (rightly, I would argue) to suppose it represents the cat before you are completely irrelevant to its representation content, according to CT. What matters, according to CT, are the causal relations between the representation and the property of being a cat; the intrinsic structural features of the representation are irrelevant. NTDP says nothing in your head has a chance of standing in the requisite causal relation to the property of being a cat unless you have a theory of cats. Lacking such a theory, nothing in your head can represent that property, perceptually or any other way.

Objection two reformulated. What you do is try to acquire a theory that will underwrite a mental representation that translates the public-language word for cats. While learning about cats, you are also learning about 'cat'. You say 'cat' whenever your detectors generate the symbol *R*. You keep refining the theory that mediates your detection system's tokening of *R* until you are saying 'cat' in the presence of the distal stimuli that prompt 'cat' in your language community.

---

[8] There is, of course, a corresponding problem for whisker detection and consequently for |whisker|s. So, perhaps, (1) should read, "*p*cats have *p*whiskers." We then work on |whisker| and |cat| simultaneously.

Reply one. This objection falsely implies that only humans (and other language users, if any) can mentally represent the property of being a cat. Cats recognize conspecifics, and hence, according to all but the most hopelessly chauvinistic version of LOT, cats represent catness.[9] Moreover, nonverbal children can recognize cats, as discrimination experiments show. Moreover, you can recognize things for which your language has no word. The objection under consideration will require that all of these be things for which you have a completely adequate paraphrase. This is surely an empirical question. Moreover, it is the sort of empirical question that a theory of content should leave open. Whatever the arguments are for LOT and CT, they should not settle issues like this.

Reply two. The objection currently on the table implies that the acquisition of 'cat' and |cat| are coeval. But is it an hypothesis dear to the hearts of LOTers that the acquisition of 'cat' is a matter of figuring out which mental representation best translates 'cat'. According to this view, learning the meaning of 'cat' is a matter of confirming the hypothesis that 'cat' means |cat| (that is, confirming '|"cat" means cat|'), and this is evidently ruled out by the scenario under consideration. Since I am not fond of the translation theory of language acquisition myself, I regard this as merely a polemical point, and I include it here only to sow the seeds of doubt in the opposition.

II. STAGE TWO: CT ENTAILS LOT

CT entails that mental representation is arbitrary in the following sense: any primitive representation could mean anything. You get this feature in CT for the following reason: if you can build a *P*-detector at all, you can arrange for it to output any symbol you like when it detects *P*.[10] So far as CT is concerned, then, any primitive symbol could mean any detectable property.

CT also entails that there are finitely many semantically primitive representations in the scheme of mental representation. To see this, begin with the observation that CT requires a detector for each primitive representation. Detectors do not have to be componential or functional units. All that is required is that, for each primitive representation, there is a mechanism/process that detects its target property. Since a finite system cannot incorporate infinitely many such mechanisms/processes, the question reduces to this: Can one mechanism/process detect infinitely many target properties?

---

[9] Maybe cats could get by with a |same species/variety as me|. Maybe. But (1) there is evidence that |me| might be problematic for cats; and (2) in the context of the larger argument, acquiring a theory that will underwrite a |same species/variety as me| is going to be just as problematic as acquiring a theory that will underwrite |cat|.

[10] It is tempting to suppose that the state of the detector when it detects *P* is the symbol in question, but this cannot be right, for then every time the system tokens that symbol, it will think it has detected *P*, and that will rule out thoughts like this: "Well, it is not a cat, but if it were, it would pounce on that mouse."

It might seem so. Consider a planar array of photo-sensors. If we assume that a square pattern, regardless of location, size, or orientation on the array, is a representation of squareness, and make a comparable assumption about every shape, then this system will (depending on density, and discounting "viewing angles" other than those normal to the array) detect and represent infinitely many shapes. But notice that CT could not possibly underwrite such a representational scheme. For what we have done here is assume that any square pattern represents squareness regardless of its regular causes. To see this, notice that a particular pattern will only be caused by some squares, namely, those with the right location, orientation, and size-distance combination relative to the array. The causal theory will say that the pattern represents only those squares, not squareness. To fix this problem, we shall have to map each square pattern onto some single arbitrary symbol. To do that, we shall have to introduce an algorithm that abstracts from size, orientation, and location on the array. It will have to look for a closed pattern with four straight sides. To distinguish square patterns from trapezoids and parallelograms, it will have to test for right-angled corners; to weed out rectangles, it will have to test for side equality. It will, in short, have to know about squares. Ditto for any other shape that is represented by a primitive in the system. Since there is only a finite amount of memory available, this knowledge will have to be finite. So, the system can only map a finite number of shapes onto primitive representations.

It is clear that this argument generalizes. NTDP requires, in effect, a theory of *P* for the detection of *P*, and CT requires the detection of *P* as a precondition of the primitive representation of *P*. In a finite system, there is only so much room for theory, so it follows that there can, at any given time, be only a finite number of primitives in play. CT allows for a scheme in which the set of primitives is unbounded in the sense that, given enough memory, another can always be added.[11] But it does not allow for a scheme in which the number of primitives is infinite.

This may seem a pretty trivial result, but other stories about content do not have this consequence. As the shape example shows, if a representation represents everything structurally isomorphic to it,[12] then it is possible to have a scheme with as many primitive shape representations as there are shapes (or something isomorphic to them) for the representations themselves to have. Finite precision will make it impossible for any actual system to exploit all that primitive representational power, but the scheme has it for all that.

---

[11] Even this may be allowing too much, since, as every personal-computer owner knows, the architecture typically limits how much memory you can add.

Thus, CT implies that the scheme of mental representation has a finite number of arbitrary primitives. Moreover, since the scheme of mental representation needs to be productive, it follows that there must be a way of combining the semantically arbitrary primitives into complex representations such that the meaning of the complex is a function of the meaning of the constituents and their mode of combination. A scheme of finitely many semantically arbitrary primitives which allows for an unbounded set of semantically distinct complex representations whose contents are determined by their constituents and mode of combination is LOT as near as makes no difference.[13] So CT entails LOT.[14]

III. CONCLUSION

I have argued that, taken together, CT and LOT are incompatible with the nontransducibility of distal properties; and I have argued that CT entails LOT. It follows that CT is incompatible with the nontransducibility of distal properties. Since distal properties are not transducible, it follows that CT is false. Since the argument uses only a generic form of CT, it follows that CT is false in all its forms.

The conclusion can be generalized a bit. Nothing in the argument given here depends on the content-fixing relation's being causal; any covariationist semantics will face exactly the same problem, since (1) NTDP says that the only way to get reliable covariation between a mental state and a distal property is to have a theory of that property, and (2) covariational accounts make representation arbitrary in the sense required by the argument of stage two.

In *Representations, Targets, and Attitudes* (*op. cit.*), I argue that CT is a restricted form of functional-role semantics. CT is what you get when you restrict the functional roles that determine content to the roles a representation plays in detection. The question therefore naturally arises as to whether functional-role semantics generally is incompatible with NTDP. I do not know the answer to that question. What is clear, however, is that nothing like the argument just rehearsed will work against functional-role semantics generally, since that argument turns essentially on facts about the detection of distal properties, and detection has no privileged role to play in functional-role theories of mental content.

---

[12] See my *Representations, Targets, and Attitudes* (Cambridge: MIT, 1996).

[13] There is nothing in CT to guarantee that the combinatorics have to be Tarskian, but, then, neither must advocates of LOT insist on Tarskian combinatorics. It is just that Tarskian combinatorics are the only ones currently available for combining arbitrary symbols productively. LOT is Tarskian by default, not by inner necessity.

[14] Strictly speaking, what I have argued is that CT plus finite-detection resources plus productivity entails LOT. The finiteness of detection resources is, I suppose, an empirical premise, but it is uncontroversial. Productivity is an empirical premise as well, but it is a pretty secure premise, and it is a premise that both CTers and LOTers typically accept.

ROBERT CUMMINS

University of Arizona

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 11, NOVEMBER 1997

QUINE'S LIMITED NATURALISM

The key components of W. V. Quine's philosophical position, taken as a whole, seem to fall into two broad categories: the austere or hard component, and the more open or soft component. On the hard side, we find commitments to:

physicalism in ontology[1]
extensionalism in logic
naturalism in epistemology
behaviorism in psychology[2]

I refer to this as the *austere* or *hard* side of Quine's position because it places strict and not easily satisfied limits on the form that an acceptable philosophical treatment of a subject can take.

CALCUTTA UNIVERSITY

On what I call the open or soft side of Quine's position, we find commitments to:

holism of a Duhemian kind
a strong doctrine of revisability

And, most famously:

the three "ins":

---

[1] Since Quine claims to be indifferent with respect to ontology, it may seem inappropriate to assign any ontological position to him. I shall return to this point later.

[2] It is not clear whether Quine is generally committed to behaviorism in psychology. He is, however, explicitly committed to behaviorism in his account of language acquisition. In *Pursuit of Truth* (Cambridge: Harvard, 1982, revised edition), he is perfectly clear on this point: "In psychology one may or may not be a behaviorist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on overt behavior in observable situations" (pp. 37-38).

0022-362X/97/9411/543–63

indeterminacy of translation
inscrutability of reference
infacticity

I call this side of Quine's position *open* or *soft* because of its commitment to a wide range of indeterminacies. Such a commitment considerably mitigates the strict demands found in the austere or hard side of Quine's position; for, if correct, it provides a principled reason for not undertaking strong reductionist programs of the kind often pursued—usually with small success—by traditional empiricists. These indeterminacies can also be used polemically to dismiss a wide range of competing philosophical positions that rest—explicitly or implicitly—on strong commitments to determinacy.[3]

If we look separately at these two clusters of commitments, the items in each seem, intuitively at least, to suit one another. Quine exhibits great skill in his writings in showing how the themes in each cluster interrelate with one another. In contrast, if we compare the two clusters as wholes, they seem, again intuitively, to display fundamentally different tendencies. Depending on which way we proceed, it can seem that all the toughness in the first cluster is given away by the softness in the second, or, going the other direction, that the openness of the second cluster can seem to be throttled by the narrow constraints of the first. Quine, of course, thinks that he has brought these apparently competing viewpoints into proper balance, a unity in difference: Bertrand Russell's logic and F. H. Bradley's logic enter the peaceable kingdom together. I think this apparent reconciliation of seemingly competing viewpoints is what gives Quine's position much of its attractiveness and resilience. The question is whether he pulls it off—more specifically, whether he pulls it off within a thorough-going naturalistic framework.

We can first look at Quine's commitment to physicalism. There is, of course, nothing in the soft cluster that commits one, or even inclines one, to physicalism. To the contrary, the doctrine of empirical underdetermination—and the various "ins" it generates—suggests that physicalism can be no more than one possible ontology among others. This was Quine's explicit view in his early philosophical essay, "On What There Is"[4]:

---

[3] For example, in *Word and Object* (Cambridge: MIT, 1960), Quine invokes indeterminacy to reject C. S. Peirce's convergence theory of truth in these words: "Peirce was tempted to define truth outright in terms of scientific method, as the ideal theory which is approached as a limit when the (supposed) canons of scientific method are used unceasingly on continuing experience.... [But] we have no reason to suppose that man's surface irritations even unto eternity admit of any one systematization that is scientifically better or simpler than all possible others" (p. 23).

> I [have] advanced an explicit standard whereby to decide what the ontological commitments of a theory are. But the question what ontology actually to adopt still stands open, and the obvious counsel is tolerance and an experimental spirit. Let us by all means see how much of the physicalistic conceptual scheme can be reduced to a phenomenalistic one; still, physics also naturally demands pursuing, irreducible *in toto* though it be. Let us see how, or to what degree, natural science may be rendered independent of platonistic mathematics; but let us also pursue mathematics and delve into its platonistic foundations (*ibid.*, p. 19).

By 1953, Quine had abandoned this generous attitude toward phenomenalism and had adopted, in its place, a commitment to physicalism, whole-hearted except for the grudging admission of a few, seemingly unavoidable, abstract entities.[5] What led Quine to alter his position in this fundamental way? This is not an easy question to answer; for, surprisingly, specific and detailed criticisms of phenomenalism are scarce in Quine's writings. In *Word and Object*, Quine makes the general claim that sense data "are neither adequate in lieu of physical objects nor helpful in addition to them" (*op. cit.*, p. 239). The first claim, that sense data are not adequate *in lieu of* physical objects, is briefly defended in the opening four pages of *Word and Object*, where Quine argues that sense-data language is dependent on physical-object language, not the other way around. The second claim, that sense data are not helpful *in addition to* physical objects, is aimed at those who think that illusions force us to introduce perceptual intermediaries, such as sense data. In response, Quine claims that illusions do not force us to introduce sense data into our ontology, because talk about illusions or illusory properties might be adequately dealt with "by a propositional-attitude construction in which 'seems that' or the like is made to govern a subsidiary sentence about physical objects" (*op. cit.*, p. 235). In fact, none of these considerations is decisive against phenomenalism or even comes close to being decisive. The idea that sense-data language is parasitic on thing-language has been around for a while, but it is far from obvious that arguments of this kind carry the day, as those familiar with H. P. Grice's[6] causal theory of perception will know. The appeal to propositional-attitude constructions to deal with perceptual illusions is no less problematic. Propositional-attitude constructions are not, after all, free on demand in Quine's program. Before they can be admitted to standing, they must be given proper extensionalist laundering, and it still remains unclear whether this can be done in a satisfactory manner. So it seems that, if Quine rejected phenomenalism on the basis of the arguments presented in *Word and Object*, then he actually had no compelling reason for doing so.

---

[4] In *From a Logical Point of View* (Cambridge: Harvard, 1953).

[5] Roger Gibson cites Quine's "On Mental Entities" as the turning point—*The Philosophy of W. V. Quine* (Tampa: Florida University Presses, 1982), p. 56.

[6] "The Causal Theory of Perception," *Proceedings of the Aristotelian Society*, Supplementary Volume xxxv (1961): 121-52.

In all likelihood, however, Quine's repudiation of phenomenalism did not rest primarily on arguments of this kind. If it did, we would expect them to be developed in more detail—including a sensitivity to possible rejoinders. Quine's reasons for rejecting phenomenalism were probably of a different kind, taking something like the following form: the contest between various competing ontological positions has been held and physicalism has won—almost hands down. Physical science has proved an unsurpassed engine of prediction and control. In contrast, phenomenalist programs, despite the heroic efforts made in their behalf, have invariably ended in failure. So even if it remains an abstract possibility that some other (nonphysicalist) standpoint could produce an equal measure of prediction and control, it would be idle (an uninteresting legalism) to insist on the point. As matters now stand, as philosophers, we begin with physicalism; it is not our task to justify it.[7]

An important feature of the response that I have put into Quine's mouth is that it involves a commitment to a quite specific conception of empirical inquiry. As far as I know, this view of inquiry first appears in a fully developed form in the writings of C. S. Peirce. It is the view that proper inquiry arises from and responds to problems that emerge at a specific stage of empirical investigation. We come on the scene armed with past achievements (funded knowledge) and hampered by past mistakes (persisting error). We are presented with a specific difficulty and our task, then, is to use, extend, or modify our past commitments in order to resolve the difficulty that confronts us. An important feature of this approach is that it treats inquiry as a historical—or at least a temporal—phenomenon. The options we examine are almost wholly determined by our past achievements. At times, a recalcitrant problem may force us to modify previous commitments, but when this happens, we do so grudgingly, trying to preserve as much as possible. Echoing these Peircean ideas, Quine speaks of conservatism or, more mellifluously, of the "maxim of minimum mutilation."[8]

---

[7] Quine makes this point in a number of places: for example, in *Roots of Reference* (LaSalle, Ill.: Open Court, 1974), where he says that the epistemologist "no longer dreams of a first philosophy, firmer than science, on which science can be based..." (p. 3).

A further feature of this Peircean conception of inquiry is its impatience with idle, purely speculative, that is, Cartesian doubts. In his early essay, "Some Consequences of Four Incapacities," Peirce[9] put it this way:

> We cannot begin with complete doubt. We must begin with all the prejudices which we actually have when we enter upon the study of philosophy. These prejudices are not to be dispelled by a maxim, for they are things which it does not occur to us can be questioned.... A person may, it is true, in the course of his studies, find reason to doubt what he began by believing; but in that case he doubts because he has a positive reason for it, and not on account of the Cartesian maxim. Let us not pretend to doubt in philosophy what we do not doubt in our hearts (*ibid.*, pp. 156-57).

"The Fixation of Belief" contains a similar remark:

> The mere putting of a proposition into the interrogative form does not stimulate the mind to any struggle after belief. There must be a real and living doubt, and without this all discussion is idle (*ibid.*, p. 232).

Quine makes the same point when he tells us that "Cartesian doubt is not the way to begin."[10] I am not saying that Quine owes his views about the nature of proper inquiry to Peirce. Perhaps he does, but where Quine could have referred to Peirce, he refers to Otto Neurath instead. My claim is that Quine, however he came by it, is committed to a view of inquiry very similar to Peirce's and, furthermore, that this is a *fundamental* feature of his philosophy.[11]

A video panel with Quine on his epistemology provided an opportunity to explore this idea.[12] Martin Davies asked Quine how he would deal with the Cartesian problem raised by the possibility that he (Quine) might, at the very moment, be asleep and dreaming. The exchange continued as follows:

---

[8] *Pursuit of Truth*, p. 14.

[9] In *Collected Papers of Charles Sanders Peirce* (Cambridge: Harvard, 1931), Volume 5.

[10] "The Nature of Natural Knowledge," in Samuel Guttenplan, ed., *Mind and Language* (New York: Oxford, 1975), pp. 67-81, here p. 68.

[11] There are other important similarities between Peirce's philosophy and Quine's, but here I am only concerned with their similar views concerning the nature of inquiry. I am not, of course, denying the existence of important differences on other matters.

[12] The videotapes of these discussions have been published by R. V. Fara under the general title, *In Conversation W. V. Quine* (London: Philosophy International, 1994). The discussion of Quine's epistemology is entitled *The Fogelin Panel*.

> Quine: I am ruling the dream hypothesis out in the sense that I dismiss it as very unlikely. And I think that is the mood in which we do our thinking generally, that there is plenty to worry about in the way of things that could interfere with our hypotheses, and show that we are wrong in them. And we worry about the likeliest ones [that is, those most likely to give us trouble].
> Fogelin: Is it fair to say that your answer is grounded in a kind of theory of inquiry...[a theory of] what honest inquiry looks like?
> Quine: Very good, yes.
> Fogelin: And these [Cartesian] doubts simply don't have a place in honest inquiry.
> Quine: Yes.

Quine's commitment to a Peircean conception of inquiry is not always evident, for it is embedded in his commitment to furthering the program of naturalizing epistemology. The following passage, which appears in the opening two pages of "The Nature of Natural Knowledge," is a model of Quine, the naturalized epistemologist, at work:

> Scepticism is an offshoot of science. The basis for scepticism is the awareness of illusion, the discovery that we must not always believe our eyes. Scepticism battens on mirages, on seemingly bent sticks in water, on rainbows, after-images, double images, dreams. But in what sense are these illusions? In the sense that they seem to be material objects which they in fact are not. Illusions are illusions only relative to a prior acceptance of genuine bodies with which to contrast them....
>
> Rudimentary physical science, that is, common sense about bodies, is thus needed as a springboard for scepticism. It contributes the needed notion of a distinction between reality and illusion, and that is not all. It also discerns regularities of bodily behavior which are indispensable to that distinction. The sceptic's example of the seemingly bent stick owes its force to our knowledge that sticks do not bend by immersion; and his examples of mirages, after-images, dreams, and the rest are similarly parasitic upon positive science, however primitive.
>
> I am not accusing the sceptic of begging the question. He is quite within his rights in assuming science in order to refute science; this, if carried out, would be a straightforward argument by *reductio ad absurdum*. I am only making the point that sceptical doubts are scientific doubts (*op. cit.*, pp. 67-68).

But are skeptical doubts always scientific doubts? In particular, can so-called skeptical scenarios be treated as scientific doubts that must be resolved, if they can be resolved, by an appeal to science? Quine was challenged on just this point in the panel discussion: he was asked to produce a scientific argument that would show that he was not, at that very moment, a brain-in-the-vat with all his experiences fed into him by a clever neurophysiologist. He took the challenge head on:

> Quine: My attitude towards [this] is the same as my attitude towards more commonplace situations where the thing is not quite so unthinkable from a naive point of view. Namely, I would think in terms of naturalistic plausibility. What we know, or what we firmly believe,...is that it would really be an implausible achievement, at this stage anyway, to rig up such a brain. And so I don't think I am one (*op. cit.*).

Setting aside its stunning straight-forwardness—which I find hard not to admire—the remarkable feature of this passage is its seeming obliviousness to the very point of the skeptical challenge. The skeptical scenario is, of course, presented as a global challenge to empirical knowledge, and even if it is not exactly question begging to offer an empirical response to it, it does, however, seem beside the point in not addressing the issue that is obviously being raised. My suggestion is that the source of this seeming obtuseness is not obtuseness at all but, instead, a deep commitment on Quine's part to a Peircean theory of inquiry. From the Peircean perspective, the Cartesian challenge is itself beside the point. I believe that Quine really meant it when he said "Cartesian doubt is not the way to begin."

I think that there is something deeply important in Quine's refusal to rise to the philosophical bait by saying more than that the circumstances presented in skeptical scenarios are empirically implausible. This is one respect in which Donald Davidson's[13] instincts are fundamentally at odds with Quine's. In some ways, Quine's response is reminiscent of G. E. Moore's[14] attempt to prove the existence of material objects simply by exhibiting one hand, then another. There is, however, a fundamental difference between Moore's response to skepticism and Quine's. Moore[15] often claimed to know, and know with certainty, things that he could not support by arguments. In contrast, Quine deals with probabilities—not certainties—and, more importantly, takes it as his task to supply supporting arguments intended to show that skeptical hypotheses simply have a low probability of being true.

---

[13] "A Coherence Theory of Truth and Knowledge," in Ernest Lepore, ed., *Truth and Interpretation* (Cambridge: Blackwell, 1986), pp. 307-19, and "Epistemology Externalized," *Dialectica*, XLV, 2-3 (1991): 191-202. In the first essay, Davidson uses what might be called a supercharged version of the principle of charity in an attempt to rule out the possibility of massive error of the kind envisaged in skeptical scenarios. In the second essay, Davidson invokes an externalist account of reference in an attempt to rule out the possibility that the references of the words that a person is using are massively skewed—as they seem to be in brain-in-the-vat scenarios. Elsewhere, I have maintained that neither of Davidson's arguments refutes historically significant versions of skepticism—*Pyrrhonian Reflections on Knowledge and Justification* (New York: Oxford, 1994). However that might be, the point here is that Quine does not employ arguments of this seemingly transcendental kind in his own responses to skepticism.

[14] "Proof of an External World," in *Philosophical Papers* (London: Allen and Unwin, 1959), pp. 126-48.

I think that Quine's way of responding to skeptical challenges makes it evident that he is committed to something very like a Peircean conception of inquiry in epistemology. I now want to ask whether he is similarly committed to such a conception of inquiry in all dimensions of his philosophical position, in particular whether he is so committed in his treatment of meaning and reference. I was prompted—though only later—to consider this issue after reflecting on a response Quine gave on the video panel when pressed on his commitment to revisability—one of the central components of what I have called the open or soft side of his position. In "Two Dogmas of Empiricism," Quine[16] famously claimed:

> Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system.... Conversely, by the same token, no statement is immune to revision (*ibid.*, p. 43).

When asked if this is not hard to take, especially when applied to laws of logic, he replied in the video discussion as follows:

> Well, I think I rather agree. I think nowadays it seems to me at best an uninteresting legalism (*op. cit.*).

To call one of his central doctrines—one of the doctrines that gave his position much of its novelty—an "uninteresting legalism" seems to represent a fundamental change in position. If so, what has produced it? My suggestion is that Quine's attitude toward the possibility that we might have to abandon fundamental laws of logic is precisely the same as his attitude toward the possibility that phenomenalism may serve as a viable alternative to physicalism. For Quine, both possibilities should be dismissed as uninteresting legalisms because they play no significant role in honest inquiry. Neither makes the short list of candidates for serious consideration.

In fact, as the center of gravity of Quine's position shifts toward naturalized epistemology, all the components in what I have called the soft cluster come under pressure. Consider his commitment to holism. In "Two Dogmas of Empiricism," he declared that "the unit of empirical significance is the whole of science" (*op. cit.*, p. 42). Taken literally, rather than hyperbolically, this makes mutual understanding among common folk very difficult to comprehend. Perhaps for reasons something like this, over time Quine has retreated considerably from his previous full-bore holism. In "Five Milestones of Empiricism,"[17] the commitment to holism is very qualified indeed:

---

[15] See "A Defence of Common Sense," in *Philosophical Papers*, pp. 32-59.

[16] In *From a Logical Point of View*, pp. 20–46.

> It is an uninteresting legalism [that phrase again]...to think of our scientific system of the world as involved *en bloc* in every prediction. More modest chunks suffice, and so may be ascribed their independent empirical meaning, nearly enough, since some vagueness in meaning must be allowed for in any event (*ibid.*, p. 71).

Yet even if Quine is now willing to say that the scientific system of the world is "very disjointed," some vestiges of his earlier commitment to overarching holism remain. In "Five Milestones of Empiricism," he tells us that:

> All sciences interlock to some extent; they share a common logic and generally some common part of mathematics, even when nothing else (*ibid.*, p. 71).[18]

It is, however, hard to imagine a much thinner form of unity. Both physicists and foresters employ the same patterns of logical inference. We can, if we like, conjoin or disjoin statements about leptons with statements about trees. If we bring mathematics in, we can count both leptons and trees. Yet it is hard to see how any of this provides a significant connection between the subject matters of these two chunks of theory. If this logical/mathematical cohesion is all that is left of Quine's earlier robust holism, then there is hardly anything left to it at all.

It seems, then, that both revisability and holism, maintained in any significantly strong sense, have both become "uninteresting legalisms." We can turn next to the signature feature of Quine's position—his commitment to various indeterminacies—and ask how they have fared as the center of gravity of his position has shifted toward naturalized epistemology. We can begin with his notion of indeterminacy of translation. No aspect of Quine's position has attracted more attention and, with it, more controversy. Rather like Ludwig Wittgenstein's private-language argument, it seems to produce conversion experiences in those who claim to plumb its depths; others just do not get it.

---

[17] In *Theories and Things* (Cambridge: Harvard, 1981), pp 67-72. In the video, Quine spoke of a "chunk of theory of a critical mass." This is rather an odd metaphor, since things blow apart when they reach critical mass—but never mind.

[18] In the video panel discussion, he remarked that "Logic applies across the board," adding: "Logic in the strict sense, as what is ordinarily formulated in the first-order predicate calculus, is shared by all of our beliefs, all of our science."

Part of the problem in dealing with Quine's notion of indeterminacy of translation is that it is not altogether easy to see what the doctrine amounts to and what it is supposed to establish. Quine has, in fact, recently said that his views on the indeterminacy of translation have been widely misconstrued. According to Quine, readers have tended to treat his discussion of the observation sentence 'Gavagai!' as concerning indeterminacy of translation when, in fact, it concerns indeterminacy (or inscrutability) of reference. In *Pursuit of Truth*, he makes the point this way:

> Ironically, indeterminacy of translation in the strong sense was not what I coined the word [Gavagai] to illustrate. It did not illustrate that, for 'Gavagai' is an observation sentence, firmly translatable holophrastically as '(Lo, a) rabbit'. But this translation is insufficient to fix the reference of 'gavagai' as a term; that was the point of the example. It is an extreme example of the indeterminacy of reference... (*op. cit.*, p. 51).

It is an important feature of Quine's position that we do not encounter indeterminacy of translation with observation sentences taken holophrastically, that is, taken as semantically unsegmented wholes. Indeterminacy of translation breaks out only when we attempt to form what Quine calls "analytical hypotheses" concerning the internal structure of sentences.[19] So the observation sentence 'Gavagai!', while exhibiting indeterminacy of reference, does not exhibit indeterminacy of translation. I shall return to the notion of indeterminacy of reference below.

Having separated indeterminacy of reference from indeterminacy of translation, we can now ask what Quine had in mind by the second notion. In *From Stimulus to Science*, Quine[20] explains the underlying idea in these words:

> What...if two field linguists spent long years going bilingual in the same remote valley while scrupulously shunning each other? Would they agree, in the end, on sameness of meaning between native standing sentences and English ones? Or suppose they both compiled manuals of translation, and both manuals proved successful in translating some

---

[19] This indeterminacy becomes progressively more radical as we move from extensional idioms to intentional idioms.

[20] Cambridge: Harvard, 1995.

> long native monologue into coherent English. Then suppose we translate it again using the two manuals alternately, sentence by sentence. Would the result still be coherent? If not, and given no basis for saying which manual is at fault, we have what I have called indeterminacy of translation (*ibid.*, p. 82).

We can represent these two translations as follows:

Translation 1: $S1_1,\ldots, S1_n$
Translation 2: $S2_1,\ldots, S2_n$

We are then invited to produce a new translation by interleaving sentences from the two translations in the following way: take the first sentence from the first translation, the second from the second, the third from the first, the fourth from the second, and so on.

Translation 3: $S1_1, S2_2, S1_3, S2_4,\ldots$

Quine asks whether this new translation would still be coherent, telling us that, if it is not coherent and there is no basis for finding fault with at least one of the manuals, then we have an instance of what Quine calls indeterminacy of translation.

The force of this argument will, of course, crucially turn on the kind of incoherence this interleaving produces. We can take it for granted that Quine has in mind more than mere stylistic incompatibility that would result if the first translator had a penchant for compound sentences whereas the second did not. An incoherence deeper than this is needed to support the strong conclusions that Quine wishes to derive from his doctrine of the indeterminacy of translation. What is needed is a clear example of the kind of incoherence Quine has in mind when radical translation is applied to a natural language. I do not think that Quine has anywhere satisfied this need and, in fact, he candidly admits in his recent writings that he is not in a position to do so:

> [T]he full or holophrastic indeterminacy of translation draws too broadly on a language to admit of factual illustration.[21]

But if the indeterminacy of translation draws too broadly on a language even to admit of factual illustration, what role can it play in Quine's avowedly naturalistic program? The answer, though Quine seems not to recognize this, is that it can have no role in such a program. To see this, we need only examine the kinds of arguments that Quine uses when attacking the idea of propositions as sentence meanings.

---

[21] *Pursuit of Truth*, p. 50. The reference to *holophrastic* is crucial. I shall return to it.

The doctrine that "propositions as sentence meanings is untenable" goes back to "Two Dogmas of Empiricism." The first move in the argument is to draw a very un-Fregean conclusion from a Fregean distinction:

> Once the theory of meaning is sharply separated from the theory of reference, it is a short step to recognizing as the primary business of the theory of meaning simply the synonymy of linguistic forms and the analyticity of statements; *meanings themselves, as obscure intermediary entities, may well be abandoned.*[22]

In "Two Dogmas," Quine's argument turned on the claim that there is no theoretically satisfactory way of explicating the notion of synonymy, since each explication involves a notion equally in need of theoretical explication. In this way, the theoretical significance of synonymy (hence sentence meaning) is undermined. In his later writings, Quine uses the notion of indeterminacy of translation to reach a similar, though more radical, result. He begins by employing a key move already used in "Two Dogmas of Empiricism":

> The meaning of a sentence of one language is what it shares with its translations in another language...

He then continues:

> so I propounded my thought experiment of radical translation. It led to the negative conclusion, a thesis of indeterminacy of translation.[23]

As we have seen, Quine's argument in behalf of this negative conclusion is that linguistic data (in principle, at least) can never foreclose the possibility that alternative translation manuals might be produced, both empirically adequate yet incompatible with each other in the sense that sentences translated using one manual could not always be coherently interleaved with sentences translated using the other. On the basis of the bare possibility that this (rather weak) form of incoherence could emerge, Quine draws his strong conclusion concerning the indeterminacy of translation.

Anyone fully committed to a naturalistic program in philosophy will find this uncongenial. From a naturalistic standpoint—which contains what I have called the *principle of honest inquiry*—it is sufficient to note that problems posed by incompatible manuals of translation do not, in fact, arise. Setting aside nuances and differences in specialized vocabularies (considerations on which Quine does not rely), translations do take place, sentence by sentence, from one language to another without Quine's problem of indeterminacy arising. As we have seen, he admits as much. Thus, to the field linguist, Quine's worry about indeterminacy of translation is an idle concern—or, to echo Quine's own phrase, "an uninteresting legalism."

---

[22] *From a Logical Point of View*, p. 22 (emphasis added). This is a crucial move on Quine's part and deserves close examination—but I shall not do that here.

[23] *Pursuit of Truth*, p. 37.

I think that Quine's naturalism, if pursued in an unrestricted manner, would also bring severe pressure against the second indeterminacy listed above: his deep commitment to the indeterminacy (or inscrutability) of reference. It would at least cast it in a very different light. The roots of the idea of the indeterminacy of reference are found in Quine's early essay "On What There Is," where he attempted to "analyze away" those paradigms of transparently referring expressions: proper names. This program of defusing ontology, begun with respect to proper names in "On What There Is," was brought to full expression in "Ontological Relativity." It has not been abandoned by Quine since.

Although the conclusion is different, Quine's strategy with respect to the indeterminacy of reference is similar to his strategy with respect to the indeterminacy of translation. Again, we are presented with a possibility of conflicting alternatives, each equally empirically supported, then asked to provide a basis for choosing between them. Our inability to do so is again used to draw a strong philosophical conclusion. Quine attempts to make his case for the indeterminacy of reference in various ways, most famously in the story he tells about the reference of the word 'gavagai', which can be taken to be a rabbit, a collection of undetached rabbit parts, a singleton containing only this rabbit as its sole member, and so on. In his more recent writings, he employs what he calls *proxy functions* as a way, to his mind, of putting the thesis of indeterminacy of reference beyond doubt. Proxy functions, as Quine defines them, are functions which are "one-to-one reinterpretations of objective reference. They leave the truth values of the sentences undisturbed."[14] Quine has given a number of examples of proxy functions, but the most elegant—and the one that seems to please him most—involves what he calls *cosmic complements*. The cosmic complement of Quine, for example, is that very large object that contains the whole universe except for Quine. It is the whole universe except for a hole where Quine is. Turning to his favorite target of reference, rabbits, Quine explains his notion of a cosmic complement in these words:

---

[14] *From Stimulus to Science*, p. 72.

> The word 'rabbit' would now denote not each rabbit but the cosmic complement of each, and the predicate 'furry' would now denote not each furry thing but the cosmic complement of each. Saying that rabbits are furry would thus be reinterpreted as saying that complements-of-rabbits are complements-of-furry things, with 'complements-of-rabbits' and 'complements-of-furry' seen as atomic predicates. The two sentences are obviously equivalent (*ibid.*, pp. 71-72).

If they do not seem obviously equivalent, look at it this way: "If all rabbits are furry things, then anything that is the cosmic complement of a rabbit is also, quite trivially, a cosmic complement of a furry thing."

With this argument in hand, Quine quickly draws a strong philosophical conclusion:

> I conclude from [the indeterminacy of reference as shown by cosmic complements] that what matters for any objects, concrete or abstract, is not what they are but what they contribute to our overall theory of the world as neutral nodes in its logical structure (*ibid.*, pp. 74-75).

As with his treatment of indeterminacy of translation, I think that we should first be struck by both the abstractness and the artificiality of the discussion. Imagine trying to pursue science using cosmic complements. Cosmic complements of ordinary objects are very strange, huge, ungainly, and disconnected entities. That they could serve a useful function in organizing scientific inquiry seems wholly remote. In fact, it seems that the only way to make the cosmic complements of ordinary objects serviceable is to speak of *their* cosmic complements, thereby, for example, getting Quine back from cosmic-complement Quine. Again, we seem to be dealing with an issue that will not arise in honest inquiry, so it is hard to understand why Quine is worrying about it at all. He acknowledges as much in *Pursuit of Truth*:

> Bodies were our primordial reifications, rooted in innate perceptual similarities. It would be gratuitous to swap them for proxies; the point was just that one could (*op. cit.*, p. 34).

But what precisely is the point of the point that we could make such a swap? It seems that, from the standpoint of a naturalized philosophy, this bare possibility is yet another uninteresting legalism.

The following passage from "Ontological Relativity"[23] is similarly perplexing:

---

[23] In *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 26-68.

> An actual field linguist would of course be sensible enough to equate "gavagai" with "rabbit," dismissing such perverse alternatives as "undetached rabbit part" and "rabbit stage" out of hand.... The implicit maxim guiding his choice of "rabbit," and similar choices for other native words, is that an enduring and relatively homogenous object, moving as a whole against a contrasting background, is a likely reference for a short expression. If he were to become conscious of this maxim, he might celebrate it as one of the linguistic universals, or traits of all languages.... But he would be wrong; the maxim is his own imposition, toward settling what is objectively indeterminate. It is a very sensible imposition, and I would recommend no other. But I am making a philosophical point (*ibid.*, p. 34).

Perhaps the field linguist would be wrong in treating his maxim as one of the linguistic universals, but to a philosopher with naturalistic instincts, the point is not philosophical but empirical, and cannot be settled by merely reflecting on the bare possibility of alternative ontological commitments.

But is Quine not making precisely the point that I claim he is ignoring? Is he not simply saying that the objects we take seriously are those which further scientific investigation? The answer to this is "no": Quine is making a much stronger claim, namely, that objects are "neutral nodes in [the] logical structure" of our "overall theory of the world." This is a thesis concerning the notion of reification and is much stronger than the commonplace that some things are theoretically more interesting than others. Thus, Quine owes us an argument going beyond the truism:

> The world contains all sorts of objects; which ones we take seriously is a function of our theoretical concerns.

to the much stronger claim:

> Objects are posits (reifications, fictions); and which posits we make is a function of our theoretical concerns.

The first claim seems unassailable, but we hardly need the notion of proxy functions to establish it. In contrast, the second thesis puts forward a complex claim concerning the operations of our cognitive apparatus. Perhaps an argument could be developed establishing its truth or at least its plausibility—I am not denying that—but it is hard to see how this could be done invoking nothing more than the abstract possibility of constructing proxy functions. After all, the ability to construct proxy functions, while establishing what we might call *notational relativity*, does not, by itself, establish ontological relativity. That the same thing can be referred to in different ways settles nothing about its ontological status.

It should be acknowledged that Quine's appeal to proxy functions might serve the useful function of dispelling the illusion, if one is under it, that logic itself forces determinate ontological commitments on us. But such an argument cannot be used to show objects are nothing more than reifications or posits. An argument to support this much stronger claim will have to come from someplace else. To illustrate what I have in mind, we can consider a simpler case. Someone might think that the existence of proper names or other singular terms in our language forces us to admit into our ontology a class of particulars to serve as referents of these referring expressions. The underlying argument, simply stated, might look like this: "Since we are stuck with proper names, we are stuck with their referents as well." In "On What There Is," Quine rejected this line of reasoning by showing how proper names are eliminable in first-order quantificational logic. Thus, if one holds—as Quine has held—that first-order quantification theory exhausts the domain of logic properly considered, then we arrive at the result that employing proper names in logical notation forces on us no ontological commitments not present when proper names are eliminated. If correct, this would be a useful result, but even if correct it would have no tendency to show that a certain class of *entities* is eliminable from our ontology as well. An independent argument is needed to do this. In "On What There Is," Quine does present such an independent argument. He holds that we cannot treat proper names as directly referring expressions because, if we do, sentences like 'Prester John never existed' will present insuperable difficulties. In particular, if the use of the name 'Prester John' commits us to the existence of Prester John, then in saying 'Prester John never existed' we would be committing ourselves to the existence of the very thing whose existence we are denying. Of course, in the age of the new theory of reference, all this is very controversial, but, if correct, it will give us what we are looking for—a good reason for going beyond the claim:

> Logic does not force us to introduce entities as referents of proper names.

to the stronger claim:

> We may (or perhaps must) exclude from our ontology entities whose purpose is to serve as the referents of proper names.

Whether Quine's treatment of proper names is correct or not, at least it has the correct form. If correct, it would do two things: (i) provide us with a reason for excluding from our ontology entities whose purpose is to serve as the referents of proper names, and (ii) show us, through the introduction of appropriate translational devices, that the elimination of proper names engenders no loss of logical power.

Does Quine have a similar two-part argument connecting the possibility of constructing proxy functions to his thesis of ontological relativity? For the most part, Quine's practice is to move directly from the possibility of constructing proxy functions to the thesis of ontological relativity and then, from there, to various ideas associated with it—for example, the indifference of ontology. It is possible, however, to find passages that suggest that other commitments are at work lubricating this transition from proxy functions to ontological relativity. The following comes from Quine's "Reply to Stroud"[26]:

> Thus imagine an inclusive theory of the world, regimented in the framework of predicate logic. Let '*Fxy*' stand for some open sentence that determines *x* uniquely for each value of '*y*' and vice versa. Now where '*P*' is a one-place predicate, reinterpret '*Px*' for each value of '*x*' as '*(Ex)(Fxy . Py)*'. Reinterpret every primitive one-place predicate in this way, and every primitive many-place predicate correspondingly. The structure of our theory of the world will remain unchanged. *Even its links to observational evidence will remain undisturbed, for the observation sentences are conditioned holophrastically to stimulations, irrespective of any reshuffling of objective reference. Nothing detectable has happened.* Save the structure and you save all.
>
> Russell urged the all-importance of structure in his *Analysis of Matter*, and Ramsey made the point more rigorously with his Ramsey sentences. Russell and Ramsey were urging the indifference only of theoretical objects, as against observational ones. *Once we take observation sentences holophrastically, however, reference and objects generally go theoretical. The indifference or inscrutability of ontology comes to apply across the board* (*ibid.*, p. 473; emphasis added).

This passage is interesting in the following respect: it begins by providing a recipe for eliminating primitive properties and primitive relations using proxy functions. Then, moving more slowly than he sometimes does, Quine goes on to draw the modest conclusion that with the elimination of primitive properties and primitive relations, "the structure of our theory of the world will remain unchanged." This establishes what above I called *notational* relativity, but so far Quine has not asserted the thesis of *ontological* relativity and the doctrine of reification associated with it. Quine announces the thesis of ontological relativity only after introducing very strong claims concerning the central role of observation sentences taken holophrastically:

---

[26] *Midwest Studies in Philosophy*, Volume VI, P. French, T. Uehling, Jr., H. Wettstein, eds. (Minneapolis: Minnesota UP, 1981), pp. 473-75. Christopher Hookway called my attention to this essay.

> Observation sentences are conditioned holophrastically to stimulations, irrespective of any reshuffling of objective reference.

and

> Once we take observation sentences holophrastically, however, reference and objects generally go theoretical. The indifference or inscrutability of ontology comes to apply across the board.

The key idea here is not (or not simply) that a system and its transformation under a proxy function will be observationally invariant. What is centrally important for matters of ontology is the reference to observation sentences being taken *holophrastically*. To take a sentence holophrastically is to treat it as an unsegmented whole. Taken holophrastically, the sentence 'The cat chased the rat' does not contain the expression 'rat' as an independent unit of meaning any more than, to use the standard example, the expression 'Socrates' does. The theoretically important aspect of sentences taken holophrastically is that, containing no internal syntactical or semantic structure, they contain no distinction between expressions' being used referentially and their being used predicatively. For this reason, observation sentences taken holophrastically cannot be otherwise than indifferent to ontology.

We now seem to have a candidate for taking us from mere notational relativity to ontological relativity. If observation sentences taken holophrastically are, as Quine says, the alpha of the scientific enterprise, then the starting point—the supplier of all empirical content—will be indifferent to ontology. It may then seem plausible to hold that ontological commitments are introduced by us—not by experience—in constructing a scientific picture of the world. In *Pursuit of Truth*, Quine puts it this way: "We began with the monolithic sentence, not the term. A lesson of proxy functions is that our ontology, like grammar, is part of our own conceptual contribution to our theory of the world. Man proposes; the world disposes, but only by holophrastic yes-or-no verdicts on the observation sentences that embody man's predictions" (*op. cit.*, p. 36). This, or something very close to it, seems to be Quine's underlying argument for ontological relativity.

There are three things to be said at this point. First, Quine's thesis of ontological relativity no longer appears as an independent doctrine that sets the stage for the project of a naturalized epistemology; for the thesis itself is now seen to rest upon the possibility of carrying out the project of constructing science on the basis of observation sentences taken holophrastically.

Second, it is far from clear that, in the various places where he has attempted to do so, this enterprise has been successful.[27] The problematic area—the sticking point, as it seems to me—concerns the transition from observation sentences taken holophrastically to sentences articulated by a subject/predicate structure. The transition is clearly crucial to Quine's project, for without it, no subject term presents itself for pronominalization, and then the project of reification (relativized or not) has no foothold. Yet when we examine the passages concerning this crucial matter, we find that they are invariably both brief and highly metaphorical. For example, in explaining the difference between the mere conjunction of the two observation sentences, 'Lo, a pebble' and 'Lo, blue', and the predication of blue to a pebble, Quine tells us that the "predication focuses the two fulfillments, requiring them to coincide or amply overlap. The blue must encompass the pebble."[28] Not only are the concepts of things *overlapping* and *encompassing* bad metaphors for predication in general, they are themselves predicative notions and thus take for granted the very thing that needs explaining. Although I cannot go into this here, I do not think that any of Quine's attempts to explicate the notion of predication in contrast with conjunction fare any better. If that is correct, then his program of passing from observation sentences taken holophrastically to a picture of the world bogs down almost at once.

Finally, it is worth noting that Quine's attempt to trace a path from stimulation to science is carried out with almost no concern for the psychological reality of the process, that is, with almost no references to empirical investigations of how language is actually acquired. The reason for this, I suggest, is that Quine's program of naturalizing epistemology is not driven by the actual exigencies of producing an empirical account of how human beings, given the cognitive capacities they possess, are able to construct a reasonably adequate picture of the universe they inhabit. Quine's inspiration comes from the library, not the laboratory. Specifically, the structure of Quine's program is set by the structure of Rudolf Carnap's program in *The Logical Structure of the World*.[29] In *Stimulation to Science*, Quine freely acknowledges this.

---

[27] For example, in *Word and Object*, briefly in *Pursuit of Truth*, and most recently in *From Stimulation to Science*.

[28] *Pursuit of Truth*, p. 4.

[29] Rolf A. George, trans. (Berkeley: California UP, 1967); translation of *Der logische Aufbau der Welt* (Berlin: Weltkreis, 1928).

> [Naturalism is the] rational reconstruction of the individual's and/or the race's actual acquisition of a responsible theory of the external world.... Unlike the old epistemologists, we seek no firmer basis for science than science itself.... It is a matter, as always in science, of tackling one problem with the help of our answers to others (*op. cit.*, p. 16).

So far so good, but then this:

> Despite this radical shift in orientation and goal, we can imitate the phenomenalistic groundwork of Carnap's *Aufbau* in our new setting (*op. cit.*, p. 16).

In contrast to this approach, a genuinely naturalized program in epistemology and philosophy of language would not begin with a prior commitment to producing a rational reconstruction along the structural lines of Carnap's *Aufbau*. It would begin with problems forced upon us by empirical inquiry itself, not with an attempt to resuscitate a failed *philosophical* program. It is important, then, to see how close Quine's program remains to Carnap's program in the *Aufbau*. Having made what he takes to be a radical shift in starting point—beginning with observation sentences taken holophrastically, rather than beginning with reports of immediate sensa—things move, as Quine acknowledges, along the structural lines of the *Aufbau*. Launched from a starting point of debatable empirical significance—observation sentences taken holophrastically—Quine's program proceeds with almost no contact with the empirical checkpoints central to doing philosophy in a naturalistic spirit.

The aspects of Quine's position I have challenged as being at odds with his naturalistic program all come from what I have called the open or soft side of his theory. We have noted that Quine, over time, has muted two components of this cluster. He has come to consider both holism and revisability, in their strong forms, "uninteresting legalisms." As we have seen, Quine also acknowledges that "the full or holophrastic indeterminacy of translation draws too broadly on a language to admit of factual illustration,"[8] so it is hard to see how this matter can play a significant role in a philosophy pursued in a naturalistic spirit. One area where Quine has been unswerving, however, is in his commitment to the related notions of ontological relativity, ontological indeterminacy, and the indifference of ontology. I have tried to show that arguments Quine uses to establish ontological relativity are ill-suited to anyone committed to a thoroughgoing naturalism in philosophy.

---

[8] *Pursuit of Truth*, p. 50.

A surprising feature of many of Quine's arguments is that they employ something very like Cartesian skeptical scenarios—though applied to issues concerning meaning and reference rather than to knowledge. We are asked to entertain possibilities having no prior naturalistic plausibility—two linguists returning from the field bearing incompatible manuals of translation, or a world segmented into the cosmic counterparts of ordinary things—and then we are challenged to find some definitive way of ruling out these possibilities. When it becomes evident that they cannot be ruled out, Quine draws strong philosophical conclusions from our inability to do so.

Of course, Quine's arguments will have force only if it is against the rules to dismiss these possibilities on the grounds that they do not present live issues for empirical investigation. In that case, we are not allowed to do unto Quine as Quine has done unto Descartes. We are not, that is, allowed to employ the naturalistic program, with its commitment to the Peircean notion of honest inquiry, across the board to include not only theory of knowledge but theory of meaning as well. We want to know why not.

ROBERT J. FOGELIN

Dartmouth College

MANIFEST KINDS*

You finally meet the love of your life. After an exciting series of dates, he decides to propose. There you are together at the restaurant, Lespinasse, enjoying a simply ravishing meal. Just before the desert is served he produces a small box, the sort of box that could very well contain an engagement ring. You open it with anticipation, only to find inside a nugget of gold and some chimney soot. Your heart sinks. You knew he was a philosopher, but you realize with growing alarm exactly what this means as he offers the following curious speech:

> Well, er...I have just been reading Saul Kripke's *Naming and Necessity*[1] in which it is decisively argued that water and H_2O are one and the same thing, numerically identical. Now this is not an idiosyncratic claim about water, it's the consequence of a general view that each manifest kind of stuff—coffee, molybdenite, plastic—is identical with the chemical kind which makes up the relevant samples of the manifest kind. So not only is the kind water numerically identical with H_2O, but so is the kind ice, the kind snow, and the kind water vapor. Therefore, by applying the transitivity of identity, we philosophers have discovered that snow is numerically identical with water vapor. But now since the point is perfectly general it applies alike to water, plastic and even diamond. The kind diamond is numerically identical to the chemical kind carbon, and so of course is soot. So diamond is identical with soot. That means you have no complaint on the score of there being no diamond. There is diamond because there is soot, for diamond is identical with soot.

You consider whether its worth pointing out that what you expected was not just diamond, as if carbonado, a sort of diamond slag, might have done. You expected a cut diamond, a precious stone, and some soot is definitely not a precious stone. But you glumly decide that you are being snowed, and that nothing could penetrate this mumbo jumbo. You suspect that, if you insisted on a cut diamond, he might well produce a piece of chiseled coal. After all, since he thinks that diamond = soot, he probably also thinks that dia-

---

* Thanks to Paul Benacerraf, David Hilbert, Saul Kripke, and Scott Soames for helpful discussions of these matters.

[1] Cambridge: Harvard, 1980; see especially pp. 99, 116, 128-29, 134, 148, 150. A characteristic remark is: "Let's consider how this applies to the types of identity statements expressing scientific discoveries that I talked about before—say, that water is H_2O" (p. 128).

0022-362X/97/9411/564–83

mond = coal, and if diamond = coal then a diamond or a piece of diamond is going to be hard for him to distinguish from a piece of coal. Despite your steely grimace, your would-be husband continues:

> No doubt you are wondering about the ring. Well now that we philosophers have established the identity of water vapor and snow, on the one hand, and soot and diamond, on the other, we are in a position to argue for the identity of a ring and the gold that makes it up. After all, the only argument against this identity was that a ring is necessarily ring-shaped while no quantity of gold is necessarily ring-shaped. I'm sure you can now see that this is a bad argument, for if it were not a bad argument then we could likewise argue that water vapor is not numerically identical with snow, because water vapor is essentially a vapor and snow is not. But to admit that water vapor is not numerically identical with snow is, by parity of reasoning, to admit that water is not numerically identical with H_2O. For if water is identical with H_2O then so is water vapor and snow. So you see why we must reject the argument that distinguishes a ring and the quantity of gold which makes it up.

You have had enough. As you leave him in the restaurant, you point to the check, saying: "Please take care of that aggregate of organic molecules."

I

Your would-be husband is a bit of a nerd, but his little sophism is quite instructive. The conclusion he aimed for—namely, that water vapor is snow and diamond is soot—is, of course, absurd. But the absurdity can be validly derived from (2), (3), and (4) below along with (1), the most commonly cited philosophical example of an identity linking a *manifest kind*, that is, a kind whose instances we identify and re-identify on the basis of their manifest properties, with a *chemical kind*, that is, a kind individuated by its perhaps recherche chemical properties.

(1) Water = H_2O.
(2) If water = H_2O then water vapor = H_2O and snow = H_2O.
(3) If water = H_2O then diamond = C (carbon).
(4) If diamond = C then soot = C.

yield:

(5) Water vapor = snow and diamond = soot.

Why is (5) absurd? Well, water vapor is a kind of vapor whose instances—fogs, mists, jets of steam, and so on—are quantities of vapor, whereas snow is a kind of powdery material whose instances are quantities of powder. Diamond is a kind of stuff whose instances are stones and lumps (of diamond), whereas soot is a kind of powdery material whose instances are quantities of (sooty) powder. If kinds are to turn out to be one and the same, then so must their instances. But a quantity of powder cannot be identical with a fog or a mist or a jet, nor can a stone or lump be identical with a quantity of powder.

So why is the little sophism at all instructive? Surely, we can account for the falsity of (2) and (3) in a way that does not reflect badly on (1). No; here I aim to show that this can*not* be done. The right account of why the argument is a sophism requires that we understand the relation between manifest kinds like water and chemical kinds like H_2O to be constitution rather than identity. It is (1) that is false.[2]

II

A first thought on behalf of the friends of (1) is this. Instances of chemical kinds come in different states or conditions: solids, liquids, gases, powders, plasmas, and so on. On the other hand, some ordinary kind names like water vapor, snow, diamond, and soot are somehow more demanding as to the condition or state of the things in question. So we might say that water vapor is H_2O in a vaporous condition; snow is H_2O in a powdery condition; diamond is C in a (largish-scale, clear, well-formed) crystalline condition; and soot is C in a powdery condition. By contrast, water can be understood as an inclusive kind: something can be an instance of that kind by being water vapor or by being snow or by being ice or by being liquid water.

Accordingly, (1) should be replaced by:

(6) The inclusive kind water = H_2O in a liquid or solid or powdery or vaporous condition, whereas liquid water = H_2O in a liquid condition.

Likewise, (2) and (3) are to be replaced by:

(7) Water vapor = H_2O in a vaporous condition and snow = H_2O in a powdery condition.

and:

(8) Diamond = C in a crystalline condition and soot = C in a powdery condition.

From all of which, we hope, nothing untoward follows. This account suggests why (2) and (3) are bogus steps. They omit any reference to the respective states crucially associated with water vapor, snow, and diamond.

---

[2] Some may already think that (1) is a misleading way of expressing an identification of water and H_2O. If you think that, feel free to take a peek at section VIII where an alternative way of parsing this identification is found wanting.

But how are we to understand a phrase like 'H_2O in a vaporous condition'? In particular, is this true?

(9) H_2O = H_2O in a vaporous condition.

If so, we have our reductio ad absurdum back again, for if (9) is true then so is:

(10) H_2O = H_2O in a powdery condition.

Thus, it follows that:

(11) H_2O in a vaporous condition = H_2O in a powdery condition.

and hence, by hypothesis, that water vapor = snow.

Clearly, we only advance if (9) is false. (I am inclined to think that (9) is true, at least on its most natural reading. After all, holy water is water and occupied Paris was Paris. But let us try to advance anyway.) What then is H_2O in a vaporous condition, and what is its relation to H_2O *sans phrase*?

III

Let us tailor the reading of 'H_2O in a vaporous condition' to suit. We want:

(9) H_2O = H_2O in a vaporous condition.

to be false and:

(7) Water vapor = H_2O in a vaporous condition and snow = H_2O in a powdery condition.

to be true. There is another constraint. Just as a diamond is essentially crystalline, a vapor is essentially vaporous. If a mist should liquefy or solidify, the mist is no more; just as when a diamond is completely crushed so that its crystalline form is destroyed, the diamond is no more. (See the remarks below about the yellow fog.)

Our tailored understanding of 'H_2O in a vaporous condition' thus needs to imply that, if something is an instance of H_2O in a vaporous condition, then it is essentially in a vaporous condition. On this understanding of 'H_2O in a vaporous condition', (7) is plausibly taken to be true, so long as we treat 'H_2O in a powdery condition' in the same way. (9) is certainly false and, indeed, it now follows that water vapor is *not* identical with snow. Once again, (2) and (3) were the bogus steps on this diagnosis.

We now have what the friends of (1) want. The cost is supposing that quite generally there is a way of forming complex names out of simpler names and qualifying phrases where the condition described in the phrase enters into the essence of the thing picked out by the complex name. In the sequel, we shall see that this is a severe cost which in fact purchases nothing in the way of a principled resistance to kind constitution.

IV

We may now say something useful about the relation between H_2O and H_2O in a vaporous condition, so understood. Let us begin with this: Is an instance of the second kind, that is, a particular vapor made up of H_2O, accidentally or contingently an instance of that kind? Could that very vapor—that particular fog or mist—have been or come to be made up of say H_2SO_4? No, it seems that a particular vapor's chemical constitution is part of what makes it that very vapor. (This, of course, is entirely compatible with supposing that other vapors are made up of H_2SO_4 or, indeed, of the philosophers' compound *XYZ*.)

Next, and more important for what follows: Could that very vapor have been a solid or a liquid? I do not mean to ask whether the stuff that makes up the vapor comes in solid or liquid form. I know that it does. I mean to ask whether, say, T. S. Eliot's yellow fog which "rubbed its back against the window pane" could have been or could come to be identical with a block of ice or a puddle of water. I do not mean to ask whether the fog could solidify into a block of ice or liquefy into a puddle of water. I know that it could. I mean to ask whether the fog could *be* either of these things. It seems to me that the fog could not be either of these things. For when the fog solidifies or liquefies, that is the end of the fog. A vapor like the fog is essentially a vapor. This fact is not at all at odds with the fact that some liquid might be the result of a vapor liquefying and that some ice might be the result of a vapor solidifying. Iron oxide is the result of iron rusting, but it is not iron. A corpse is the result of killing a living human being, but it is not that living human being.

To summarize: the instances of H_2O in a vaporous condition, that is, the water vapors, are essentially instances of that kind. Now, whenever we have a water vapor we have a quantity of H_2O, that is, some instance of the kind H_2O. But obviously the very same quantity of H_2O—say, a particular mole of H_2O—can be first in the form of a vapor, then in the form of a liquid, and, finally, in the form of a solid. So, whereas a water vapor is essentially vaporous, the quantity of H_2O that makes it up is not essentially vaporous. Therefore, no water vapor is identical with the quantity of H_2O that makes it up. The relation in question is material constitution, not numerical identity. When the instances of one kind—for example, H_2O—in this way materially constitute the instances of another kind—for example, H_2O in a vaporous condition—I shall mark the intimate relation between the kinds by saying that the first kind constitutes the second.

V

Where are we? The friend of:

(1) Water = H_2O.

can only defend it by maintaining that there are kinds of stuff denoted by such designators as 'H_2O in a vaporous condition', kinds distinct from the kinds picked out by the simpler names they contain. Of such kinds they then must say things like:

(12) H_2O in a vaporous condition is not identical with but is constituted by H_2O.

meaning that every instance of the first kind is constituted by but not identical with an instance of the second kind.

The bump in the carpet has just been moved, rather than removed. For it is now entirely unclear why at the very beginning we should not have captured the relevant fact of chemistry not by (1) but by:

(13) Water is constituted by H_2O.

Or better, given the facts about heavy water and tritium,[3] by:

(14) Water (indeed pure water) is constituted by a mixture in which H_2O is by far the largest component.

Taking this rather than (1) to be the content of the relevant chemical discovery also avoids having to suppose that there are such distinct kinds as H_2O at 72°F and H_2O at 72.00003...recurring° F and so on and so forth. This endless multiplication of kinds is the consequence of the special construal of 'H_2O in a vaporous condition' forced on the friends of (1). For in order to defend that special construal, they need to claim that quite generally there is a way of forming complex names out of simpler names and qualifying phrases where the condition described in the phrase enters into the essence of the thing picked out by the complex name.

---

[3] In 1932, the chemist Harold Clayton Urey discovered that water contains a small amount (1 part in 6,000) of so-called heavy water or deuterium oxide (D_2O)—deuterium being the isotope of hydrogen with an atomic weight of 2. In 1951, Aristid Grosse discovered that naturally occurring water also contains minute traces of tritium oxide—tritium being the isotope of hydrogen with an atomic weight of 3. Most manifest kinds of stuff are mixtures.

The fact that naturally occurring water contains tritium shows that it is *chemically* impure but not that it is impure in any ordinary sense. Pure water could still be a mixture of H_2O, deuterium oxide, and tritium, with H_2O the massively dominant ingredient. Compare air: Who would suppose that argon is an impurity in the air? For more on the restricted application of the notion of *chemical* purity, see the discussion of gold below, particularly footnote nine.

Better to drop that special construal and treat the relevant designators as functioning like 'occupied Paris'. Then we do not have a distinction between kinds for every distinction we can capture by any qualifying phrase. Accordingly, 'H_2O at 72° F' and 'H_2O at 72.00003...recurring° F' both pick out the kind H_2O. That is the other advantage of recognizing the constitution of kinds where others see only identity. We avoid an endless and implausible multiplication of kinds.[4]

VI

I used a form of argument, at the end of section IV, which is not respected in some quarters. I argued against identifying a particular vapor with the quantity of H_2O that makes it up, on the grounds that the vapor is essentially vaporous—for example, ceases to exist when solidified into ice—while the same quantity of H_2O could be first in a vaporous form and then in a solid form. Some philosophers attempt to resist this kind of argument by finding a disguised relativity in modal predication. A characteristic form of resistance would here amount to this: the vapor qua vapor is essentially vaporous, but so also is the H_2O qua vapor. Elsewhere, I have tried to discredit this whole line of thought (*ibid.*). Here, I simply wish to point out that the locus classicus for the claim that water = H_2O implicitly repudiates this line of thought. For in *Naming and Necessity*, Kripke argues against the identity of a statue and the matter that makes it up on the grounds that the statue but not the matter is essentially statuesque.[5]

My point is simply that the very same consideration applies in the case of kinds of stuff. Kinds are identical only if they have just the same instances. Modal arguments precisely paralleling the argument about the statue and its matter apply to the instances of manifest kinds and the instances of the chemical kinds that make them up.

---

[4] Compare the argument of my "Constitution Is Not Identity," *Mind*, XCI (1992): 83-102, where it is shown that those who identify a ring with the gold that makes it up will find it hard to resist an implausible result, namely, an endless multiplication of overlapping rings.

[5] See *Naming and Necessity*, footnote 74, and the discussion of persons and their bodies in the surrounding text.

Those arguments exhibit a distinction among instances which implies a distinction among the kinds they instantiate. As against the grain of *Naming and Necessity*, we should grant the same rights to kinds of material stuff as we do to kinds of material objects. If a statue is not identical to the matter that makes it up, then an instance of a manifest kind like water is not to be identified with the instance of H_2O that makes it up. Since kinds are identical only if their instances are, ordinary manifest kinds of stuff are not to be identified with underlying chemical kinds.

VII

That would be a significant result even if it admitted of a few exceptions. In this spirit, some have suggested to me that water is a special case. On their view, because we know water in liquid, solid, vaporous, and powdery forms, when using 'water' we are picking out the kind that chemistry tells us could be in any one of these forms, that is, H_2O.[8] 'Water' carries with it no restriction as to the chemical state or condition water is in, so what it picks out can be outright identified with H_2O. On this view, (1) is true, but not (2) or (3), so that the absurdities do not follow. For while being an instance of the kind water is not demanding with respect to form, being an instance of the kind snow nevertheless requires being a powder, being an instance of the kind water vapor requires being a vapor, being an instance of the kind diamond requires being a lump or a stone and being an instance of the kind soot requires being a powder (or at least a cake of powder). Lamentably, according to this view, philosophers have once more presented the atypical case as central.

Although it concedes almost all I wish to say, this view is still too sanguine. What does the advocate of the view say about the relations among water, snow, and water vapor? He had better not say that snow = water in a powdery condition and water vapor = water in a vaporous condition. For, once again, how are we to understand the designators on the right-hand sides of these identities? The dilemma, by now familiar, is: *either* we invoke the special construal, in which case we push the bump to another point in the carpet, having to admit that water merely constitutes water in a vaporous condition, *or* we face the absurd consequence that water vapor = snow.

Neither will do. Therefore, on the view under discussion, 'water' had better be a name for the stuff that merely constitutes water vapor and snow. The view can now be restated as follows: we know water as the stuff that constitutes or makes up water vapor and snow and ice and so on, hence there is no reason not to identify it with H_2O, which chemistry tells us makes up these kinds.

---

[8] David Hilbert urged this. I thank him for his insightful comments on the topics discussed in this section.

Even this view faces a serious embarrassment. How does water so understood stand to the familiar *liquid* that fills lakes and rivers and in which we bathe? Does it merely constitute that liquid or is it identical with that liquid?

Here is a way of highlighting the embarrassment. It looks to me as though whatever reason there is for saying that a water vapor is essentially vaporous is a reason for saying that a watery liquid is essentially a liquid. After all, when a liquid is completely evaporated, the liquid is no more. The liquid does not survive as the resultant vapor.[7] A particular liquid is essentially a liquid, just as a particular fog is essentially a vapor. So the advocate of the view that water constitutes water vapor, snow, and ice had better not say that water = the kind liquid water (also known as the liquid, water); for then he should have to maintain that instances of water are all of them liquid and essentially liquid.

Why would that be a bad thing to say? Well, this latter claim is in tension with the claim that water constitutes water vapor. For together these claims entail that whenever we have, say, a jet of steam, the water that constitutes the steam is a liquid, which is clearly false. To avoid that consequence, the view under discussion had better reduce to this: 'water' is a name for the stuff that merely constitutes water vapor, snow, ice, *and* liquid water.

That is rather bizarre. Who would have thought that water merely constitutes liquid water, so that none of their instances is identical? No, surely the truth is that every instance of liquid water is an instance of the inclusive kind water. This is why the view under discussion is too sanguine and is not a plausible way to save (1). It is H_2O

---

[7] What happens if after evaporating a liquid we then liquify the resulting vapor? I am not sure which of two things to say. One is: we recover the same quantity of liquid but not the very same liquid. The other is that instances of kinds of material stuff can exhibit intermittent existence; hence we recover the very same liquid. Mutatis mutandis for solidifying a liquid and then melting it. The second seems to me slightly more plausible. Compare this: any high-end stereo system is plausibly taken to be capable of intermittent existence, for if it does not exist until its assembled from the myriad of components out there in hi-fi land, then it probably does not exist when it is disassembled for transporting or for cannibalizing for other systems. If after a round of cannibalizing of components I get back my original components and reassemble them, then that is my stereo system once again.

In any case, one thing that is really implausible is to identify some liquid with the solid ice it yields and the vapor it came from. For then we should have no choice but to identify each of these with a quantity of the underlying chemical kind. A year ago that quantity could have been scattered throughout the Pacific Ocean and the Black Sea. It follows that the ice existed then. Yet this is clearly wrong.

and not the inclusive kind water which merely constitutes liquid water. So once again the thing to say is that H_2O is not identical with water.

The simple truth about the inclusive kind water is that it is neither identical with nor does it constitute any of the more restrictive kinds snow, ice, water vapor, or liquid water. Rather, these are ***subkinds*** of the inclusive kind water. This means that each instance of these kinds is an instance of the kind water.

What about saying that each instance of snow, ice, water vapor, or liquid water is an instance of H_2O? No, that too would be false. Instances of H_2O are quantities of H_2O. One of these quantities—for example, a mole of H_2O—can first be in a powdery, then a solid, then a liquid, and finally, a vaporous form. But that is not true of any of the instances of snow, ice, water vapor, or liquid water. Their form is essential to their being the particular powders, solids, vapors, and liquids that they are. The right thing to say is that these kinds are constituted by H_2O; that is, their instances are constituted by quantities of H_2O. They are not subkinds of H_2O; for they share no instances with H_2O.

VIII

There remains one more way to defend some version of the identification of water with H_2O, a way which, though initially attractive, in fact leads nowhere other than back to the conclusion that water is not to be identified with H_2O, but is only constituted by it.

Perhaps, so this initially attractive line of thought goes, in conceiving of the claim that water is H_2O as an *identity* statement flanked by rigid designators for kinds, *Naming and Necessity* was making a little too much of the analogy between identifications involving names for individuals and identifications involving names for kinds. Accordingly, theoretical identities like water is identical with H_2O, gold is identical with the element with Au, and so on, might be rewritten in this form:

(15) For any x (x is water, if and only if x is H_2O).

Whereas the claim that water vapor is H_2O can only mean:

(16) For any x (if x is water vapor, then x is H_2O).

Likewise, the claim that snow is H_2O comes to this:

(17) For any x (if x is snow, then x is H_2O).

From which nothing absurd about the relation of water vapor and snow follows.

Obviously, since I think that water, water vapor, and snow are merely constituted by H_2O, I agree with this as a way out of the problem. For if 'x is H_2O' simply means that x is constituted by H_2O, then (15), (16), and (17) are indeed true. (Forgetting about whether *XYZ* and so forth could be other kinds that constitute water.)

But equally obviously, (15), (16), and (17), so interpreted, are much too weak to capture the force of the theoretical identifications of kinds in which Kripke was rightly so interested. Throughout the third lecture of *Naming and Necessity,* Kripke is concerned with examples like gold = Au and water = H_2O partly to bring discredit on the identity theory of mind as exemplified in such claims as pain = C-fibers firing. No identity theorist would have been satisfied with claiming that pain is merely constituted by C-fibers firing.

Is there some stronger truth in the offing that preserves the import of the identity theory? Suppose we say:

(18) For any x (x is an instance of the kind water if and only if x is an instance of the kind H_2O).

modeled on:

(19) For any x (x is an instance of the state of being in pain, if and only if x is an instance of the state of having one's C-fibers firing).

Then the relevant claims about snow and water vapor will take the form of:

(20) For any x (if x is an instance of the kind snow, then x is an instance of the kind H_2O).

and:

(21) For any x (if x is an instance of the kind water vapor, then x is an instance of the kind H_2O).

from which nothing absurd about the relation between snow and water vapor follows. That is why, so this line of thought has it, (2) and (3) were the bogus steps in the original sophism.

If this is the reader's preferred way out of the sophism with which we began, then I apologize for making him sit on his hands until now. I do have an excuse. As it happens, this is not a way to save the identification of water with H_2O, even when this is understood as amounting to (18). Claims (18), (20), and (21) are false, and for reasons we have already canvased.

Eliot's yellow fog was an instance of the kind water vapor. It was also made up of a quantity of H_2O, an instance of the kind H_2O. But contrary to (21), these instances are distinct. Suppose the fog suddenly liquefies. Then we still have the very same quantity of H_2O as made up the fog, but now it is in a liquid form. (After all, part of the point of talking of a chemical kind like H_2O is to allow that particular quantities of it can first be in one state and then another.) Not so with the fog. When the fog liquefies, the fog is no more; just as when a gold ring is rolled into a ball, the ring is no more. The fog is not identical with the quantity of H_2O that makes it up and, since that is the relevant instance of the kind H_2O here, the fog is not identical with any instance of H_2O. So Eliot's fog is a counterexample that shows that (21) is false.[8] Mutatis mutandis with snow and (20).

Now (18) can look for a moment as if it is immune to this criticism. For, after all, we are concerned in (18) with the inclusive kind water, the kind which includes as subkinds water vapor and snow and ice and liquid water. When the fog liquefies, we have some water—an instance of the inclusive kind water—around.

Let us look more closely. Something is an instance of the inclusive kind water by being an instance either of the kind water vapor or of the kind snow or of the kind ice or of the kind liquid water. The inclusive kind water is a determinable kind. Things get to be instances of it by being instances of its determinates. Compare red; things get to be red by being some specific shade of red. In the case of red, there could not be a red that is nonspecific as to shade. But maybe the inclusive kind water is in this way different. Maybe one wants to allow mere quantities of water among the instances of the kind water; quantities that can be first in a solid form, then in a liquid form, then in vaporous form, and, finally, in a powdery form. Fine; but it would be unmotivated to exclude particular fogs as also instances of the inclusive kind water. If they are instances of the kind water, then they are counterexamples to (18).

What if they are not? What if we stipulate that the inclusive kind water has as instances only quantities of water, only things that have no essential tie to any particular manifest state or condition? Grand; now (18) *is* true, but then neither snow nor water vapor nor ice nor liquid water will count as subkinds of the *inclusive* kind water. For none of their instances is an instance of that kind. This means the name 'water' so understood has then been stipulated to be just a name for H_2O, whose instances are all likewise only mere quantities. If we take refuge in this stipulation, then no thesis is then being offered about the relation between manifest kinds like water and chemical kinds like H_2O.

---

[8] No one should want to say that there are here two instances of the kind H_2O which happen to be exactly co-incident at least before the liquefaction—that is, the fog and the quantity of fog which makes it up. Or if they do they should recognize that they are already in the constitution business: on their view, oddly enough, constitution holds among instances of the very same kind, namely, H_2O. Better to say that the fog is not an extra instance of the kind H_2O.

IX

One source of the urge to identify water with H_2O comes from the idea that the designator 'water' functions semantically as if it were explicitly introduced to designate the underlying chemical kind found in demonstrated samples that may include a certain liquid, a certain vapor, a certain powder, and a certain solid. For this "as if" story to be apposite it had better not burden ordinary users of 'water' with concepts that none of them nor their "experts" could yet have.

Obviously, as it stands, the "as if" story is not apposite. The introduction of the designator 'water' in this way is a performance possible only at a comparatively late stage of human history, since the notion of a chemical kind, a kind individuated by its chemical properties, probably postdates alchemy and may even require some anticipation of the classification of elements found in the periodic table. Therefore, the semantic account associated with that model of introducing 'water' can hardly be the right account of the meaning of ancient ordinary language names for the manifest stuff which we drink, bathe in, and use to irrigate land. In abstracting away from the specific states in which we find water and fixating directly on the chemical kind, the semantic account makes 'water' too much like 'Californium' or 'Einsteinium', names explicitly introduced to denote chemical kinds.

X

There are intermediate cases, cases in which the ancients may be held to have deferred to something which turned out to be modern chemistry. Indeed, 'gold' can look like such a case. A longstanding human fixation concerns the assaying of the purity and authenticity of alleged samples of gold. Chemistry gives a definitive account of what distinguishes real gold from fool's gold, and of what gold's purity amounts to. So here, at least, there is more justice in seeing chemistry as filling out a well-recognized gap in our knowledge of how to tell real gold, thereby showing us what it is to be identical with real gold.

The case for the identity theory of gold (gold = Au) is not fully convincing, however. The ancient interest in the purity and authenticity of alleged samples of gold was not a brute or nonderivative interest in its underlying chemical structure, whatever that might turn out to be, but rather an interest in gold's enduring manifest properties of malleability, splendor, resistance to corrosion, and so on; properties that are not enduring properties of fool's gold. Compare diamond and its manifest properties of hardness, crystalline form, and decorative appeal. Even though we are concerned to distinguish real diamonds from fakes and defective stones this does not establish that diamond = C, for some of the things we wish to set aside as defective or fake are forms of ballas and bort, which are also made up of carbon.

Our interest in diamond is an interest in a manifest kind. The manifest form of diamond is crucial to being a diamond. I take something like this to be so even in the case of gold. For if by 'golden' we mean something with the manifest form of gold (which may, of course, be temporarily hidden by dirt or other encrustation), then, while for the ancient goldsmiths it was probably necessary that:

(22) Any instance of gold is golden.

it is certainly not necessary that:

(23) Any quantity of Au is golden.

For single atoms of Au have no manifest form or at very least, if electron microscopes or other exotic instruments make manifest their form, not a form which is in any way golden. They could not, for example, be splendid or malleable.

Better then to say that:

(24) Gold is Au in a sizable lattice structure.

and resist the thought that we have here the 'is' of identity by asking whether 'Au in a sizeable lattice structure' is to be understood so that:

(25) Au = Au in a sizeable lattice structure.

is true. If (25) is true, then treating (24) as an identity statement runs foul of the fact that (22) and (23) have a different status. Alternatively, if (25) is thus false, so that Au in a sizeable lattice structure is only constituted by Au, then we are without any motive for resisting the claim that gold is not identical with but is constituted by Au.

I do not expect the point about what was probably necessary for the ancient goldsmiths to itself carry the day in favor of constitution. A reinforcing consideration comes from the fact that almost all the gold we have ever encountered is a mixture of Au and other ingredients. These mixtures can vary significantly without there being any question but that what we have before us is a lump or nugget or coin of gold. Even "pure" gold like the Canadian maple leaf coin contains some traces of ingredients other than Au. A single kind despite variable chemical mixtures suggests constitution and not identity.

What then of this kind: *chemically* absolutely pure gold? Certainly, it is not the kind gold, for unlike the kind gold no instances of it were found near San Francisco or Ballarat in the second half of the last century.[9] There never was nor is there likely to ever be a chemically-absolutely-pure-gold rush.

As Bishop Butler might have put it: everything, including gold, is itself and not another thing.

XI

One side effect of chemistry telling us more and more about the chemical properties of chemical kinds is that chemistry shows us which chemical kinds *constitute* manifest kinds. It does this by showing how the chemical kinds in question account for the causal profile of (instances of) the manifest kinds in question. There is still room to allow that in special cases a chemical kind could just *be* a manifest kind. Imagine that some chemists hypothesize the existence of *Kripkenium*, a transuranic element with atomic number 290. Thanks to its theorized atomic structure, they are, we may suppose, led to speculate that Kripkenium could be stable for an hour or so under very cold conditions. Centuries later, middle-sized lumps of Kripkenium are discovered on the surface of Neptune. The name 'Kripkenium' was introduced solely to denote a chemical kind, and this very chemical kind turns out to be observed in nature some time later. Likewise, it might have been that in the case of polyethylene, a hypothesized organic polymer with a certain chemical structure first received the name, and then sheets of it were made.[10]

---

[9] At least on any ordinary understanding of 'found'. Perhaps some will say that when you find a nugget of gold with its impurities, you thereby find a quantity of Au, namely, the quantity of Au atoms which is the dominant ingredient in the nugget. Allowing such a way of speaking is harmless and does not alter the main point in the text. Indeed, it is inconsistent with the claim that gold = chemically pure Au, for it amounts to the admission that the kind gold has an instance—the nugget—only a proper part of which is an instance of the kind chemically pure Au. That is an infallible sign that the kinds are distinct.

Suppose we now ask whether Au = chemically pure Au. If it is, then since coins and nuggets, which are instances of gold, are not chemically pure then gold is not to be identified with Au. So a consequence of the identity theory of gold is that Au is a distinct kind from chemically pure Au. How then are they related? It must be that the second is a subkind of the first; that is, every instance of chemically pure Au is an instance of Au, but not vice versa. On the identity theory, Au has nuggets and coins as instances, but these are not instances of the kind chemically pure Au, which presumably only has quantities of Au atoms as instances. But then a consequence of the identity theory of gold is that whenever there is a nugget of gold, there are at least two instances of Au: the nugget that is just an instance of Au and the quantity of Au atoms which is also an instance of chemically pure Au. So let us not hear the charge of "double counting" leveled against the constitution theory of manifest kinds.

Indeed, if a nugget includes as a proper subpart an instance of chemically pure Au, namely, the quantity of atoms which partly compose it, then what is the principled ground for not counting every subquantity of this quantity an instance of chemically pure Au? Notice what this would mean on the theory that has it that gold = Au and, accordingly, that chemically pure Au is a subkind of Au: every nugget would represent an immense number of instances of *gold*.

Finally, there is the problem of *degrees* of impurity. Is there a kind chemically pure Au to degree .999 which has as instances no nuggets but some coins like the Canadian maple leaf? Here we rejoin the path from the identity theory to an implausible multiplication of kinds.

This is for me a bit of a "do not care." But as it happens, I still find reason to resist. Even if all the above were so, it will still be true that the kind lump of Kripkenium is constituted by the kind Kripkenium, and that the kind sheet of polyethylene is constituted by the molecular kind polyethylene. For even here what is manifest to us are not *mere* quantities of this or that chemical kind, as it might be moles or half moles or atoms or molecules, but rather lumps or sheets made of a quantity of the relevant chemical kind. The instances of chemical kinds are mere quantities, quantities capable of being in whatever state the chemical kind can be in, whereas the instances of manifest kinds are more intimately tied to the identity conditions of the lumps or sheets or bodies or powders or vapors that those instances come in.

Here is a speculative consideration in favor of that claim about Kripkenium and polyethylene, a consideration offered as a stimulus for further debate rather than as a crucial resting point. Consider a single atom of Kripkenium or a single molecule of polyethylene. These are quantities of, and so instances of, the relevant chemical kinds. But they are not instances of any manifest kind, for they have no manifest features.[11] So, since kinds are identical only if their instances are, then Kripkenium and polyethylene, as opposed to the kinds lump of Kripkenium and sheet of polyethylene, are not manifest kinds.[12]

---

[10] Organic polymers or plastics were first made in the 1860s, partly in response to a competition sponsored by the famous American billiard and pool ball manufacturers, Phelan and Collander. They were looking for a synthetic substitute for ivory. There is likely to be among the organic polymers several examples of molecules that were first described and then synthesized.

[11] Why should every instance of a manifest kind be manifest, that is, able to be recognized and re-identified on the basis of manifest features? The case to consider is a case where we have two kinds K and K', where K has only manifest instances and K' includes these and still other instances not themselves manifest. I say that K is the manifest kind and is a subkind of K', which is not itself a manifest kind. So far this is just stipulative. In a larger manuscript entitled *The Manifest*, however, I argue that requiring that all the instances of a manifest kind or property be

In any case, however these special examples turn out, I should sum up like this: the scientific image of the world is an image of how manifest kinds and their manifest instances are constituted so as to have the efficient causal powers that they do in fact have. Manifest kinds are thus satisfactorily located in a causal-explanatory view of the world. But, in general, they are not to be "reduced" by way of identification with chemical kinds. Their manifest form is at least as important to being what they are as is the matter that makes them up.

XII

What do we get for being the champions of constitution as against identity? I believe it changes the shape of the philosophical landscape in a number of important ways. Here, in closing, I restrict myself to one example, an example which looms large in the third lecture of *Naming and Necessity* and which has been exceedingly influential ever since.

In that lecture, Kripke articulated a seemingly important asymmetry between scientifically motivated theoretical identifications such as water = H_2O, and such physicalist identifications as pain = C-fibers firing. Following Kripke, Frank Jackson and David Chalmers[13] have recently made much of this asymmetry, to the apparent discredit of the identity theory of mind. To put the point in its simplest form, when it comes to water and H_2O then an appeal to empirical discoveries and Occam's razor provides premise (26), while the a priori semantic account of 'water' as a natural kind term yields (27):

---

manifest gives a satisfactory and explanatory restriction on the dissectiveness of manifest kinds and properties.

To put it all too briefly, the manifest property red found on surfaces is often said to present itself as dissective, tha is, such that every part of the red surface is red. At least this is the common and somewhat plausible philosophical interpretation of the manifest homogeneity of redness. Given scientific realism, however, it immediately follows that nothing is red since there are small parts of any allegedly red surface which are not red. Something has gone wrong. The way out is to note that since homogeneity is a manifest property of redness it can have only manifest instances, so that the "insensible parts" of a red surface need not themselves be red. In this way, the requirement that every instance of a manifest property be itself manifest does necessary explanatory work. Similar explanatory work is done by the corresponding constraint on kinds.

[12] Suppose that as matter of chemical law Kripkenium had to come in a middle-sized lump. Even then we should have to distinguish the kinds Kripkenium and lump of Kripkenium, for instances of the former kind would include physically undetachable atomic parts that are not instances of the latter kind.

[13] Jackson, *The John Locke Lectures* (Manuscript 1995); Chalmers, *The Conscious Mind* (New York: Oxford, 1996).

(26) At an appropriately basic physical level, H_2O is the unique explainer of the causal powers characteristic of water.

(27) If at an appropriately basic physical level H_2O is the unique explainer of the causal powers characteristic of water, then water = H_2O.

(Think of Kripke's natural-kind semantics as embodying the following "as if" story: 'water' was introduced to denote the kind that explains, at an appropriately basic physical level, the causal powers of *these*—as it turns out watery—samples.)

Now, if (27) is true, then it is true a priori, that is, thanks to semantics alone. So (26), which is a discovery of physical chemistry strengthened by an appeal to simplicity or Occam's razor, yields as an a priori matter:

(28) Water = H_2O.

That is to say that it is not conceivable that (26) is true and (28) is false. There is really no explanatory gap at all between (26) and (28), for it is a priori that if (26) then (28).

What about the corresponding argument in the case of a conscious state such as pain? Perhaps empirical discoveries in neurophysiology along with simplicity considerations yield:

(26′) At an appropriately basic physical level the firing of C-fibers is the unique explainer of the causal powers characteristic of pain.

The counterpart of (27) is (27′):

(27′) If at an appropriately basic physical level the firing of C-fibers is the unique explainer of the causal powers characteristic of pain, then the firing of C-fibers = pain.

Together they entail:

(28′) Pain = the firing of C-fibers.

It is *not* a priori, however, that if (26′) then (28′); for (27′) does not hold a priori. The semantics of 'pain' is not the semantics of a natural-kind term introduced to denote the natural kind that presents the appearance of pain. Pain just is this appearance. That is why there cannot be "fool's pain"; anything that manifestly seems like pain is pain. Here, then, is the alleged explanatory gap. Even if it is true that if (26′) then (28′), we cannot *see* how it can be true in the way we can see, that is, derive a priori, that if (26) then (28). That is why peculiarly philosophical fears about "Zombies" (who have their

C-fibers firing but no pain) cannot be put to rest. As Chalmers emphasizes, such beings cannot be ruled out a priori. There must be a further "conceptually contingent" or a posteriori connection in the case of the firing of C-fibers and pain; hence the thought that pain is an "emergent property" not captured by neurophysiology. Alex Byrne[14] independently made the same point.

Our ruminations about kind constitution have three implications which together upset this influential consensus about how to formulate a central issue concerning mind and body. The first implication is that the Kripke-Jackson-Chalmers-Byrne story cannot be the right account of the explanatory gap. For, as we have seen, neither (27) nor (27′) holds a priori; both are false. At most, H_2O constitutes water and the firing of C-fibers constitutes pain. (Recall what this means: in the relevant class of cases, every instance of the second kind is constituted by an instance of the first kind. This means that the constituting instances account for the causal profile of the constituted instances.)

The truths that replace (27) and (27′) are:

(27*) If at an appropriately basic physical level H_2O is the unique explainer of the causal powers of water, then H_2O constitutes water.

(27′*) If at an appropriately basic physical level the firing of C-fibers is the unique explainer of the causal powers of pain, then the firing of C-fibers constitutes pain.

The second implication is this: *both* of these truths are plausibly taken as a priori; that is, they follow from what we should mean by "material constitution." The constituting matter of a thing is what (at an appropriately basic physical level) accounts for its characteristic causal powers. There is no explanatory asymmetry when we have the right explananda in mind, namely, that water is constituted by H_2O and pain is constituted by the firing of C-fibers.

The third implication serves to relocate the explanatory gap, and to generalize it. It is one thing to explain exhaustively the causal powers of a manifest kind. This can be done and when it is done it establishes the relevant constitution claim: the underlying explaining kind constitutes the manifest kind. It is quite another thing to identify the qualities of the manifest kind with quantitative properties at the explaining level, or to give an a priori derivation of the qualities from the quantities. Except for some very special cases, this can never be done. I think it cannot be done for manifest shape, nor for slipperiness, nor for redness, nor for wetness. To use Jackson's[15] now famous test, in my view a test he misapplied to supposed *qualia*: you can know all the relevant quantitative information without being in a position to know what these *qualities* are like. What some call the hard part of the mind-body problem is perfectly general and has nothing much to do with special features of mind and body.

---

[14] *The Emergent Mind*, Princeton doctoral thesis (1993).

Generalizing the mind-body problem in this way may simply increase the ontological anxieties of some thinkers. Speaking for myself, it makes me calmer about the whole issue. I think: "That is just the way nature is."

MARK JOHNSTON

Princeton University

---

[15] "Epiphenomenal Qualia," *Philosophical Quarterly*, XXXII (1982): 127-36.

Private Consciences and Public Reasons. KENT GREENAWALT. New York: Oxford University Press, 1995. 225 p. Cloth $45.00, paper $19.95.

In recent years, "principles of restraint" have come to play an important role in liberal political philosophy. It is now widely believed that liberal citizens should exercise a kind of self-restraint when they make political judgments or when they discuss political issues. According to different versions of the argument, citizens should avoid appealing to comprehensive views, or controversial conceptions of the good, or beliefs that are inaccessible to their fellow citizens; rather, they have a duty to rely on shared premises and ways of reasoning when they enter the public arena. Self-restraint is clearly related to state neutrality; both are seen as responses to the cultural and religious diversity that characterizes Western democratic societies in the twentieth century.

In *Private Consciences and Public Reasons*, Kent Greenawalt examines the main arguments for and against restraint and defends what he calls an "intermediate" position. In the United States, he notes, the "central issue is most strikingly posed...by religious convictions that have political implications" (4), so it is hardly surprising that nearly all of his examples involve religious beliefs. But his conclusions are also intended to apply to secular comprehensive views, such as greatest happiness utilitarianism and "traditional Marxism."

Greenawalt argues that the "basic principles of liberal democracy do not themselves resolve whether citizens and officials should exercise self-restraint" (6). Those principles preclude efforts to suppress particular religions or burden particular religious practices. Where "religious imposition" of this sort is not involved, however, the premises of liberal democracy create a presumption against restraint: citizens should be free to rely on any beliefs they regard as valid, "*unless* a substantial reason emerges why they should not" (52). Whether this presumption can be overridden depends on a society's history, culture, present composition, and likely evolution. Inevitably, the case for restraint will be stronger in some societies (even some liberal democracies) than others.

Apart from the emphasis on historical and cultural context, Greenawalt's approach has two distinguishing features. First, he suggests that different principles may apply to people who occupy different social roles. There is no reason to assume at the outset that ordinary citizens and public officials (or, for that matter, all public

0022-362X/97/9411/584–87

officials) should be subject to the same restraints. Second, he argues that there are important differences between restraints on individual judgment and restraints on public discourse. Most people find it extremely difficult to set aside their deepest convictions when they make decisions—in part because they have no guarantee that others, including their opponents, are doing likewise. Restraints on public speech are psychologically less demanding—at least one can usually know whether others are observing them. Since Greenawalt assumes that "plausible principles must be ones that most relevant people can live up to in substantial degree most of the time" (132), it follows that restraints on judgment will be more difficult to justify than restraints on speech.

In the United States, Greenawalt believes, the argument for restraint applies most forcefully to judges. They should always cast their opinions in terms of public reasons and should generally employ such reasons in deciding cases. Ordinary citizens, on the other hand, should be free to rely on comprehensive views, whether making decisions or talking politics. Legislators, executive officials, and certain "quasi-public citizens" (for example, newspaper editors and presidents of large corporations) occupy a complicated middle ground: they may give some weight to their own comprehensive views when they consider difficult questions, but they have a duty to speak the language of public reason when discussing specific issues. Greenawalt suggests that these proposals correspond at least roughly to current practice in the United States, though they are "more nuanced than those commonly proposed" by other philosophers (7).

I want to focus on what I see as the most problematic feature of Greenawalt's position: his suggestion that legislators and other public officials should exercise self-restraint when they talk about issues. The case for restraint seems weakest when we consider how it would apply to nonreligious comprehensive views. Again, Greenawalt's examples of these include "traditional Marxism." Suppose, then, that in response to the depression of 1998, a Marxist were elected to Congress from New York City. Should we expect him to debate the merits of proposed laws without invoking the basic tenets of Marxism? To be sure, he might decide, for tactical reasons, to avoid quoting from *Capital*, or referring to surplus value, or the fetishism of commodities, in speeches on the floor of the House of Representatives. But it seems bizarre to suggest that he has a moral duty to do so.

If we confine ourselves to religious views, the case for restraint seems more plausible on its face (and probably closer to this society's shared understandings). But it is still worth asking: Why should we require legislators and other public officials to avoid expressing their own deepest convictions when they discuss issues like welfare reform, or gay marriage, or assisted suicide? In the passage that comes closest to answering this question, Greenawalt argues as follows: legislators are supposed to represent all their constituents. When they express religious views, constituents who do not share those views "are likely to feel imposed upon in the sense of being excluded." While "intense religious politics in the United States probably would not produce extensive outright violence,...we are still far from harmonious mutual respect and tolerance. Religious divisions are still very significant in many regions, and people are acutely conscious of whether they are in a majority or a minority. Given all the reasons why people find themselves living in particular places, the existence of a federal system with opportunities to move to different states is not a sufficient protection for those in a minority" (157).

Greenawalt is not making a simple, straightforward argument here. But the main idea seems to be that unrestrained discourse undermines tolerance and exacerbates religious divisions, ultimately posing a threat to stability. If this is Greenawalt's position, two preliminary points are in order. First, although stability is obviously an important value (no one wants to live in the middle of a civil war), a liberal society should probably be willing to tolerate *some* threats to it in the name of religious and political liberty. Second, experience suggests that most of us tend to exaggerate threats to stability, and such mistakes can be costly. This is not a ground for rejecting stability arguments out of hand, but it does suggest that we should approach those arguments with caution, asking in each case whether there are good reasons to believe that the alleged threat is real.

In this case, I do not believe there are good reasons. While Greenawalt refers to (and warns us not to underestimate) "the harm of a religious politics" (158), he provides little evidence to suggest that the threat to stability is serious, at least in this society. (Elsewhere, it is certainly serious. But Greenawalt understands that societies differ widely on this score—that is why the case for restraint needs to be made in terms of *our* history and culture. We are not Bosnia.) This lack of evidence is especially disconcerting given Greenawalt's insistence that the "burden of persuasion" rests with those who favor restraint.

Sometimes, in the passage quoted above and elsewhere, Greenawalt hints at a different argument, one based on fairness rather than stability. Unfortunately, this argument is never worked out in any detail, so it remains unclear why citizens are being treated unfairly—and denied equal respect—when their representatives express views they do not share. The fact that citizens feel excluded when their representatives advocate policies on religious grounds is certainly relevant to this discussion, but it does not settle the question.

A fairness argument would have the advantage of not resting on any controversial empirical claims. But it would also face one difficulty worth noting. There is a tension between the appeal to fairness as a justification for self-restraint and Greenawalt's willingness to allow public officials to base (some of) their judgments on their own religious views. If anything gives rise to claims of unfairness, it ought to be the fact that lawmakers are relying on religious doctrines when they make decisions. If Greenawalt believes that basing a decision on religious grounds is morally unobjectionable, it is hard to see how he can also hold that defending the decision in religious terms violates the principles of fairness. It is not clear that these two positions can be reconciled.

Greenawalt is at his best when analyzing the arguments of those writers who favor more extensive restraints. He sets out their views with great care and offers a variety of thoughtful criticisms. But in the end he does not make a compelling case for his own moderate principles of restraint. One finishes this book with a sense that it will not be easy to override the presumption in favor of liberty.

ROBERT AMDUR

Columbia University

Making Choices: A Recasting of Decision Theory. FREDERIC SCHICK. New York: Cambridge University Press, 1997. 163 p. Cloth $49.95; paper $14.95.

This short, elegantly written book sets itself an ambitious agenda, undertaking to offer an overview of individual decision making under certainty, under risk, when values are multidimensional, and when they are conflicted. Topics in game theory, social choice, and moral deliberation are all included. Frederic Schick does not pretend to cover the technical matters that need to be addressed in a text book. Instead, he examines deep issues that text books often do not stress. As a consequence, the book should be an excellent companion to texts used in courses on decision theory in philosophy, psychology, and social-science departments.

The most interesting aspect of this small book, however, is the elaboration and summary of themes Schick has defended in previous work. The book ought to put Schick's important contributions in perspective for contemporary philosophers of mind, morals, and decision making.

Schick's conception of the act of choice is encapsulated in the following remark:

> ...we have options, our mind is still open, we don't yet want this or that. To make a choice is to make up our mind, to settle some issue we face: To *choose* is to come to want to take this or that option we have (11).

Schick is entirely right in claiming that in the context of deliberation, the options available to the agent are open. The agent is committed to being in suspense as to which of the available options will be implemented. Given the conviction the agent has that his choices are effective,[1] the agent who chooses x is committed to becoming certain that x will be implemented upon his choosing x. So choosing involves closing one's mind, as Schick also says; that is, the agent becomes convinced that x will be implemented. The choosing coincides with a coming to believe. On his own principles as I understand them, when we make up our mind we come to believe. Yet Schick claims that making up one's mind is coming to want.

Prima facie, the reverse is true. Choosing x is an outcome of a deliberation where the agent concludes that x is the option he wants to implement. His wanting to do x is the agent's reason for choosing to

---

[1] See Schick, "Self-knowledge, Uncertainty, and Choice," *British Journal for the Philosophy of Science*, xxx (1979): 235–52.

0022-362X/97/9411/588–97

do *x*. Once *x* is chosen, the desire to do *x* ceases; for *x* is as good as done. Far from choosing *x* being tantamount to coming to want *x*, choosing *x* seems better correlated with ceasing to want *x*.

Schick seems to think (12) that because there can be a time lag between choosing and implementing, there needs to be some explanation for why the agent remains motivated to implement the choice during the interim. That motivation is the want that comes into being with choice. It is true that once the choice is implemented, the want tends to dissipate; but motivation of some kind is needed to sustain the implementation. It is that motivation, whatever it is, that constitutes the want.

Consequently, on Schick's view, wanting *x* is not a reason for choosing *x* in the causal or explanatory sense in which he is interested. Wanting *x* does not explain choosing *x* but is the state to which the agent shifts in choosing to do *x*. Wanting *x* is not a cause of choosing *x*; but choosing *x* or coming to want *x* is a cause of implementing *x*.

Sometimes, however, choices can be explained by appeal to reasons. According to a widely discussed view, explaining why the agent chose *x* appeals to premises stating that the agent believes that *x* is an action of type *y* and *x* desires or wants to take actions of type *y*. These premises may explain the agent's choosing *x*—that is, the agent's coming to want *x*. It is desire for *y* and not for *x* that appears in the belief-desire explanation-rationalization.

Schick rightly thinks this kind of explanation is defective, because the covering laws on which such explanations are based are false. Agents facing issues sometimes have the requisite beliefs and desires but fail to choose as the reason-explanation requires.

> Suppose that a rich old lady has died. Her niece is suspected of killing her, for she stood to inherit the money. The niece believed that killing her aunt would make her rich and she wanted to be rich. She had a suitable belief-and-desire, so "she had a reason." The niece had a motive, or *grounds*. Suppose, however, the butler did it. The niece's motive didn't move her, and this though there was no change of mind or failure of effort or the like (14-15).

Schick points out that the niece might not have been moved to act even though she desired, "all things considered," to become rich and believed that killing her aunt would be the best means to that end. That is to say, the niece's failure to act might have been a case of weakness of the will as construed by Donald Davidson. Davidson's view implies that the niece may be charged with the form of irrationality he calls "incontinence." The niece might be moved by a belief and a desire even though the motive is not the all-things-considered belief and desire. Schick insists that in order to have a satisfactory explanation of why one belief-desire pair is motivating rather than another, some further feature needs to be invoked.

Schick is right to recognize considerations such as this as presenting serious difficulties for belief-desire explanations of choice. One might respond to the challenge in one of two ways: (a) abandon belief-desire models for explaining choice, or (b) modify belief-desire models by introducing one or more extra factors in identifying the reason that causally explains the choice. Schick favors response (b).

Schick suggests that the third factor involved in reason explanations are *understandings*. When an agent's choice of *x* has a reason, the reason depends on a belief, desire, and an understanding. An agent can have a belief and a desire without an understanding. Such a belief-desire pair may be *grounds* for choice but not a motivating *reason* for choice. Grounds for choice become a reason for choice when they are activated in virtue of an understanding. The agent can have disconnected belief-desire pairs—grounds for choice that are not motivating as well as proper reasons where the grounds are activated.

The notion of understanding, Schick insists, is primitive. He does not reduce or analyze it in other terms. Even so, we should have a system of postulates that impose constraints on the application of the notion of understanding and its relation to other ideas. Schick offers us many examples of the way understanding functions in explaining choice, but he provides no such constraints. Without more structured constraints on the notion of understanding, it is very difficult to see how saying that certain beliefs and desires are motivating and others are not because of the way the agent understands his options does more than saying that the motivating beliefs and desires are motivating because they are motivating and are not motivating because they are not.

In all fairness, Schick makes a substantial effort to provide us with reasons for trying to rescue the notion of understanding from obscurity, even if he himself does not succeed in doing so. The argument of the book is an argument for attempting to understand understanding. If his argument is successful, Schick will have opened up avenues for philosophical reflection, even if he has not offered us a finished map of the new terrain.

Note has already been taken of one aspect of Schick's case for taking understanding seriously. If we want to rescue the belief-desire model as a schema for explaining choice, a third factor needs to be invoked in order to address the question of weakness of the will. Schick offers us several others.

Our values are often alleged to be multidimensional. In deciding whether to purchase a Honda or a Volvo, a consumer may regard his goal or desire to be to purchase a car that effects a suitable balance between cost, safety, ease of maintenance, and comfort. Hondas are cheaper and have better maintenance records than Volvos, but Volvos do better in the other two categories. The prospective purchaser of a car still faces a problem: What constitutes a "suitable balance" between the various desiderata? Schick considers the desire in this case to be ambiguous because it is some sort of aggregation of several other desiderata.

Consider the predicament of Lillian Hellman who was called upon to testify before the House Un-American Activities Committee. If she testified regarding what she knew of the activities of her friends, she would violate her principles; and if she refused to testify, she would risk going to jail. Schick considers this, too, to be decision making under ambiguity. Hellman was concerned to "balance" two competing desires.

According to Schick, Hellman had a desire to testify (to avoid jail) and a desire to refuse to testify (to adhere to principle). She believed she could not do both. According to Schick's definition, she was in a conflict (58). He thinks that, if Hellman had found a way to balance the two desires, she would not have changed the desires. They would remain what they were. Nor would her beliefs change. But by reaching some kind of balance, Hellman would have found a way partially to activate each element of the pair of desires. Hellman would have reached a sort of understanding.

But is it true that Hellman had the wants or desires Schick claims she did? Let us first look at the case of the prospective car buyer. Would it be charitable to attribute to the car buyer a desire to minimize purchase costs no matter how dangerous the car is, no matter how poor the maintenance record is, and no matter how uncomfortable it is? No intelligent car buyer would harbor such a desire. Yet such a car buyer might organize his preferences for purchasing cars so as to rank cars according to some form of weighted average of utility functions representing preference rankings that take cost minimization alone into account, safety maximization alone into account, ease of maintenance alone, comfort alone, and so on. The mere fact that an agent's preferences for his options can be represented by a weighted average of desiderata such as these does not imply that the preferences represented by the desiderata are preferences the agent now endorses or ever did endorse. If this is a fair description of the predicament of the intelligent car buyer, the intelligent car buyer is not now and never was in conflict. Perhaps, unintelligent car buyers sometimes are found in such conflict. Schick's aim was surely not to explain the choices of unintelligent actors while neglecting intelligent ones.

What about Hellman? Did she desire at any point in the scenario to adhere to principle no matter how imprudent, or to be strictly prudent no matter how drastic the violation of principle? It is true that before reaching a decision, Hellman did not rule out following prudence in evaluating the options and did not rule out following principle. But that is not adherence to either desideratum in the sense in which they are constituents in an evaluation she might come to endorse in the given setting. To believe fully a factual proposition is not the same as admitting that it might be true. Similarly, to endorse a certain way of evaluating or preferring alternatives is not the same as failing to rule it out as a permissible standard for such evaluation.

As I understand the Hellman example, she was not in any sort of conflict as specified by Schick's definition. Hellman did not want to obey both principle and prudence knowing full well she could not do both. She was in doubt as to whether to obey principle or prudence. While in doubt she did not want to obey principle and she did not want to obey prudence. Hellman was in conflict in the sense that she was in suspense between conflicting preferences. She was not in conflict in the sense that she desired options she believed she could not implement together.

To be in conflict in this second sense (the sense of Schick's definition) is to be incoherent. No doubt we are often incoherent in this way. Some authors seem to suggest that when agents have inconsistent beliefs and have to act or reach conclusions, they should separate out consistent strands of belief and select one for use. Schick seems to embrace a conception of paraconsistent desires similar to such a view of paraconsistent beliefs. We have incoherent desires but we can still make decisions. We can ground our decisions on a consistent strand in our desires. The factor determining the selection is an understanding.

Thus, Schick represents the situation of the prospective car buyer or of Hellman or any other case of multi-attribute decision making as a situation where the agent has paraconsistent and, hence, inconsistent desires. The question of which of the incompatible desires "move" the agent is answered by saying that it is determined by understandings. According to an alternative representation, the agent is in suspense between rival evaluations of the available options, each of which can be represented as a weighted average of utility functions representing a given roster of desiderata. When the doubts are sufficiently extreme, the agent may not have a reason (that is, belief and desire) for choosing one of the options rather than another. In that case, the agent may deliberate in order to relieve doubt *by eliminating utility functions from the profile of permissible utility functions and, hence, of desires* to a point where the set of options that are admissible is substantially reduced.[2]

According to Schick's approach, motivation requires appeal to an "understanding" that selects a coherent strand in the incoherent system of preferences as choice determining. The alternative view with which I am contrasting Schick's account suggests that the system of preferences need not be incoherent. The agent may be in doubt as to what he desires. To be moved to act, the doubt must be reduced.

As long as we focus on belief-desire models and modifications of them as models designed for explaining and predicting choice, I find it hard to choose between Schick's vision of multi-attribute decision making and the one I suggested. More accurately, I find it all too easy to reach a verdict. Both models ought to be rejected. To my way of thinking, we are not going to succeed in explaining behavior by appealing to propositional attitudes or, indeed, even to attenuated versions of the intensional stance unless such models are so attenuated that the intensionality has been stripped away.

Variants on belief-desire models are of more relevance when we seek to regulate the attitudes and justify choices. Schick and I can agree that agents often do find their preferences and desires to be inconsistent or incoherent. He denies that one should be obliged to remove the inconsistency in order to rationalize a choice. One can instead select a consistent strand in one's desires to use together with one's beliefs in order to justify choosing one option over others. Such justification needs, therefore, to invoke an appeal to understandings.

---

[2] As it happens, Hellman consulted Abe Fortas who suggested a new option to her. She could offer to testify about herself but no one else without pleading the Fifth Amendment. Once the new option was included in the set of options to be ranked, it became clear to Hellman that although there was still a risk of prison, it was far less than if she refused to testify at all and there was no violation of principle. So in one dimension, the option comes out as one of the best, and on the other (prudence), it came out much better than other options that conformed to principle. One could recommend that option, provided one avoided striking a balance so strongly in favor of prudence as virtually to rule consideration of principle out of court. It would turn out superior to all other options, including others conforming to principle, if one balanced very heavily in favor of principle. Given that Hellman was not totally committed to prudence over principle, it was the option to take—which she did.

This account of permissible preference rankings and indeterminate utilities was discussed at length in my *Hard Choices* (New York: Cambridge, 1986). This discussion is an elaboration on themes to be found in my "On Indeterminate Probabilities," this JOURNAL, LXXI, 13 (July 18, 1974): 391-418; and *The Enterprise of Knowledge* (Cambridge: MIT, 1980).

By way of contrast, I think it is desirable to eliminate inconsistency and incoherence. When confronted with incoherence in our views, our initial response should be to move to a position to doubt and conduct inquiries to remove the doubt.

My disagreement with Schick does not concern questions of description, prediction, and explanation. It is a dispute about whether we ought to move to a position of suspense when we find our values, preferences, or desires to be inconsistent or incoherent in order to undertake inquiry from a position of doubt.

This is not the place to attempt to settle the dispute; but it is important to appreciate that there are at least two responses to the presence of conflict in values. Schick leaves the impression that his view is uniquely sensitive to the presence of such conflict so that the apparent fact of conflict supports his position. The alternative view sketched above acknowledges that *both* conflict in the sense of inconsistency and conflict in the sense of doubt can occur in the deliberations of agents. It argues that conflict in the sense of doubt is preferable to inconsistency. Assuming that Hellman and the car purchaser are fairly reasonable, their conflicts should be interpreted so as to reflect doubt and not inconsistency.

Schick goes to great lengths to suppress the possibility of doubt or suspense with respect to desires or preferences. He acknowledges that sometimes decisions must be faced in multi-attribute decision making where the decision maker cannot fix on a precise weighting. But Schick does not think of this as a form of suspense with respect to the weighting. Rather, it is vagueness in the weighting parameter—a vagueness that the available resources for measurement cannot eliminate (45). It is important, however, to distinguish *imprecision* (or *vagueness*) in evaluating utilities, probabilities, or preferences and *indeterminacy* in such evaluations. One might be in doubt as to which of a precisely specified range of weights to use in aggregating competing desiderata without being in any doubt that one has failed to rule out as impermissible any weighting in the specified range. In such a case, the doubt concerns what is better than what and not what one thinks about the matter. By way of contrast, one might be in no doubt regarding the numerical weight but remain in doubt as to as to the weight on which one has settled. In that case, one is vague about one's determinate preference. Of course, one might be in suspense between a range of weights and so have indeterminate preferences, and also be in doubt as to what range of weights represents that indeterminacy so that one's indeterminate judgment is also vague. By conflating vagueness with indeterminacy, Schick is tacitly replacing doubt with respect to preference with self-reflective doubt as to what one's determinate desires are.

Schick's argument for understandings as the factor *X* in improving belief-desire explanations of choice takes a different turn in his discussion of the "nonextensionality" of reasons. Consider situation I where the agent is given $1,000.00 and is asked to choose between (A) getting another $500.00 for sure and (B) an even chance between getting another $1,000.00 or nothing. Contrast this with situation II where the agent is given $2,000.00 and then asked to choose (C) between $500.00 and (D) an even chance of losing $1,000.00 or nothing. O. Kahneman and A. Tversky report that most experimental subjects in situation I choose (A) and most in situation II choose (D).

Option (A) in I and (C) in II provide for receiving the same sum in the end. So do (B) in I and (D) in II. When matched in this way, the preferences are reversed. Like Kahneman and Tversky, Schick says that preferences are reversed by the way the facts are "described" or "framed."

Such preference reversal may seem to be a form of conflict or inconsistency. We may still explain choice in terms of beliefs and desires, provided the beliefs and desires are consistent strands of the inconsistent system selected by an understanding or framing. In contrast to multi-attribute decision making where Schick's approach conflates suspension of judgment with endorsement of incompatible desires, no such diagnosis seems available here.

Schick contends such preference reversal violates a certain extensionality principle. He reports two versions of the principle; but I shall focus on the version that Schick himself takes as the main target—principle *S*. As he states the principle, it runs as follows:

> ...if you believe that *x* and *y* report the same fact (same event or situation), the values you set can't differ; you have to value the fact qua *x* the same as you value the fact qua *y*. How you *understand* that fact—whether you see it as *x* or as *y*—cannot make any difference (55).

I am not clear why principle *S* so formulated poses problems for the prevalent responses to the two problems.

Why is winning $500.00 as a result of choosing option (A) in I "the same fact" as losing $500.00 as a result of choosing (C) in situation II? Presumably, because they are both cases of ending up with $1,500.00. But this is confused. *Relative to the information available in situation* I, the sentence (a) 'The decision maker wins $500.00 from option (A)' is equivalent in truth value to (b) 'The decision maker ends up with total wealth of $1,500.00 from option (A)'. *Relative to the information available in situation* II, (c) 'The decision maker loses $500.00 from option (C)' carries the same truth value as (b′) 'The decision maker ends up with total wealth of $1,500.00 from option (C)'.

Do (b) and (b′) describe the same fact or event or situation? It is not clear that they describe facts, events, or situations at all. In situation I, (B) may be chosen so that (b) does not describe a fact at all. Likewise in II, (D) may be chosen so that (b′) does not describe a fact. In any case, if situations I and II are both implemented, it is unclear to me why one would say that (b) and (b′) report the same fact even if (A) and (C) are chosen in their respective situations.

To my way of thinking, *S* had better be reformulated as follows:

> If agent *X* is in a state of full belief *K* (representable by all sentences he is committed to being certain are true), and if *K* entails that sentences *h* and *h′* carry the same truth value, then *X* should value the truth of *h* and the truth of *h′* the same.

From principle *S* so stated, it becomes mandatory that (a) in situation I is to be evaluated the same as (b) in situation I. Likewise (c) in situation II is evaluated the same as (b′) in situation II.

Nothing follows from this concerning the comparison of (b) in I with (b′) in II. Depending on contextual cues that might be available, one might conclude from the reports of Kahneman and Tversky that (b) and (b′) are not valued the same. Principle *S* does not require it. In that case, there is no inconsistency or conflict or "preference reversal."

Of course, someone might value (b) and (b′) the same and do the parallel type of evaluation for payoff sentences for options (B) in I and (D) in II so that some sort of preference reversal results. But none of the experiments involving so-called framing effects seems to require construing decision makers as committed to such reversals; and a principle of charity argues against it. In any case, the key issue here is not whether people do or do not conform to principle *S*. Sometimes they do and sometimes they do not. The question is whether they should.

Space does not permit elaborating on the topics mentioned here or entering into an examination of Schick's contrast between rationality and sociality, of game theory, and of personal identity over time. He writes provocatively and intelligently about all of these topics. His arguments and proposals should be taken very seriously—especially by those who, like myself, think his arguments and conclusions ought to be resisted. Schick's discussion should stimulate reflection on the topic of conflict in desire, value, or preference among both those familiar with the issue and newcomers to the topic.

ISAAC LEVI

Columbia University

NEW BOOKS: ANTHOLOGIES

BARNES, JONATHAN, and MIRIAM GRIFFIN, eds.: *Philosophia Togata II: Plato and Aristotle at Rome*. New York: Oxford, 1997. viii, 300 p. Cloth $78.00.

COPELAND, JACK, ed.: *Logic and Reality: Essays on the Legacy of Arthur Prior*. New York: Oxford, 1997. x, 545 p. Cloth $112.00.

CRISP, ROGER, and MICHAEL SLOTE, eds.: *Virtue Ethics*. New York: Oxford, 1997. 284 p. Cloth $65.00, Paper $17.95.

DAVIES, STEPHEN, ed.: *Art and Its Messages: Meaning, Morality, and Society*. University Park: Penn State UP, 1997. 128 p. Paper $13.95.

DEAN, JODI, ed.: *Feminism and the New Democracy: Resisting the Political*. Thousand Oaks, CA: Sage, 1997. ix, 274 p. Cloth $75.00, Paper $26.95.

DWORKIN, GERALD, ed.: *Mill's* On Liberty: *Critical Essays*. Lanham, MD: Rowman and Littlefield, 1997. xiii, 189 p. Cloth $34.00, paper $12.95.

EDWARDS, PAUL, ed.: *Immortality*. Amherst, NY: Prometheus, 1997. xii, 337 p. Paper $19.95.

GLOCK, HANS-JOHANN, ed.: *The Rise of Analytic Philosophy*. Malden, MA: Blackwell, 1997. xiv, 95 p. Paper $19.95.

HAUGELAND, JOHN, ed.: *Mind Design II: Philosophy, Psychology, Artificial Intelligence: Revised and Enlarged*. Cambridge: MIT, 1997. viii, 476 p. Paper $18.50.

HOUSER, NATHAN, DON D. ROBERTS, and JAMES VAN EVRA, eds.: *Studies in the Logic of Charles Sanders Peirce*. Bloomington: Indiana UP, 1997. xiii, 653 p. Cloth $49.95.

JOST, WALTER, and MICHAEL J. HYDE, eds.: *Rhetoric and Hermeneutics in Our Time: A Reader*. New Haven: Yale, 1997. xxiii, 406 p. Cloth $40.00.

MELLOR, D.H., and ALEX OLIVER, eds.: *Properties*. New York: Oxford, 1997. vi, 276 p. Cloth $60.00, paper $18.95.

MILLICAN, PETER, and ANDY CLARK, eds.: *Machines and Thought: The Legacy of Alan Turing, Volume I*. New York: Oxford, 1997. x, 297 p. Cloth $55.00. *Connectionism, Concepts, and Folk Psychology: The Legacy of Alan Turing, Volume II*. New York: Oxford, 1996. ix, 281 p. Cloth $55.00.

NICHOLSON, LINDA, ed: *The Second Wave: A Reader in Feminist Theory*. New York: Routledge, 1997. x, 414 p. Paper $22.95.

PASSERIN D'ENTREVES, MAURIZIO, and SEYLA BENHABIB, eds.: *Habermas and the Unfinished Project of Modernity: Critical Essays on* The Philosophical Discourse of Modernity. Cambridge: MIT, 1997. viii, 305 p. Paper $17.00.

PETERSON, GRETHE B., ed.: *The Tanner Lectures on Human Values, Volume XVIII*. Salt Lake City: Utah UP, 1997. vi, 319 p. Cloth $30.00.

PITTMAN, JOHN P., ed.: *African-American Perspectives and Philosophical Traditions*. New York: Routledge, 1997. xxii, 296 p. Cloth $65.00, Paper $18.95.

PUTNAM, RUTH ANNA, ed.: *The Cambridge Companion to William James*. New York: Cambridge, 1997. xii, 406 p. Cloth $59.95, paper $19.95.

REATH, ANDREWS, BARBARA HERMAN, and CHRISTINE M. KORSGAARD, eds.: *Reclaiming the History of Ethics: Essays for John Rawls*. New York: Cambridge, 1997. 415 p. Cloth $59.95.

ROBINSON, JENEFER, ed.: *Music and Meaning*. Ithaca: Cornell, 1997. xii, 261 p. Cloth $47.50, Paper $16.95.

SHAPIRO, IAN, and WILL KYMLICKA, eds.: *Ethnicity and Group Rights: Nomos XXXIX*. New York: University Press, 1997. xiv, 627 p. Cloth $45.00.

SKRUPSKELIS, IGNAS K., and ELIZABETH M. BERKELEY, eds.: *The Correspondence of William James: Volume V, 1878-1884*. Charlottesville: Virginia UP, 1997. lxiv, 677 p. Cloth $60.00; and *William and Henry James: Selected Letters*. Charlottesville: Virginia UP, 1997. xxxii, 570 p. Cloth $39.95.

SLATER, JOHN G., and PETER KÖLLNER, eds.: *A Fresh Look at Empiricism: The Collected Papers of Bertrand Russell, Volume X*. New York: Routledge, 1997. xxxix, 886 p. Cloth $185.00.

SMITH, BARBARA HERRNSTEIN, and ARKADY PLOTNITSKY, eds.: *Mathematics, Science, and Postclassical Theory*. Durham, NC: Duke, 1997. vi, 279 p. Cloth $49.95, Paper $16.95.

TAIT, WILLIAM W., ed.: *Early Analytic Philosophy: Frege, Russell, Wittgenstein*. Chicago: Open Court, 1997. vii, 291 p. Cloth $54.95, Paper $26.95.

TAYLOR, C.C.W., ed.: *Oxford Studies in Ancient Philosophy, Volume XIV*. New York: Oxford, 1997. 312 p. Cloth $68.00.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIV, NO. 12, DECEMBER 1997

UNIV.

WHY CONSTITUTION IS NOT IDENTITY*

Many ordinary things are made up of, or are constituted by, material things. For example, Michaelangelo's *David* is constituted by a particular piece of white marble; Mother Teresa was constituted by a particular human body; the first Union Jack was constituted by a particular piece of cloth, and so on. Exactly what this relation of material constitution is, however, has been the subject of vigorous debate.[1] Prominent philosophers[2] have claimed that the relation between a material thing and the thing that constitutes it is identity. In contrast to such philosophers, I want to resuscitate an essentialist argument against the view that constitution is identity. The form of argument[3] I shall defend is this:

---

* I am indebted to Fred Feldman and to Max Cresswell for comments on a draft, and to Katherine Sonderegger for advice on presentation.

[1] For a useful taxonomy of solutions to famous puzzles about constitution, see Michael C. Rea, "The Problem of Material Constitution," *Philosophical Review*, CIV, 4 (October 1995): 525-52. One of the assumptions that generates the puzzles is what Rea calls *the identity assumption*: $\forall x \forall y \forall ps[(\text{the } ps \text{ compose } x \text{ at } t \ \& \text{ the } ps \text{ compose } y \text{ at } t) \rightarrow (x = y)]$. This thesis is also called *mereological extensionality*, and, as Rea notes, it is often expressed by the claim "constitution is identity" (p. 528). Assuming that persons and bodies, say, are wholly composed of exactly the same molecules, my article may be seen as an attack on mereological extensionality, understood as Rea's identity assumption. For further discussion, see Peter Simons, *Parts: A Study in Ontology* (New York: Oxford, 1987).

[2] See, for example, Allan Gibbard, "Contingent Identity," *Journal of Philosophical Logic*, IV (1975): 187-221; Anil Gupta, *The Logic of Common Nouns* (New Haven: Yale, 1980); David Lewis, "Counterparts of Persons and Their Bodies," this JOURNAL, LXVIII, 7 (April 8, 1971): 203-11; Denis Robinson, "Re-identifying Matter," *Philosophical Review*, XCI, 3 (July 1982): 317-42; Stephen Yablo, "Identity, Essence, and Indiscernibility," this JOURNAL, LXXXIV, 6 (June 1987): 293-314.

[3] If instances of this argument form are sound, then two things can occupy the same place at the same time. Friends of that view (though not necessarily friends of my argument that entails it) include Simons; David Wiggins, *Sameness and Substance* (Cambridge: Blackwell, 1980); Frederick C. Doepke, "Spatially Coinciding Objects," *Ratio*, XXIV, 1 (1982): 45-60; Judith Jarvis Thomson, "Parthood and Identity

0022-362X/97/9412/599–621

(1) x is essentially an F
(2) y is not essentially an F
$\therefore$ (3) $x \neq y$

Arguments of this form are widely thought to be question begging, by both essentialists and anti-essentialists.[4] My plan is to set out a particular instance of an argument of the form (1)-(3) and to defend it from both anti-essentialist and essentialist challenges. If we use the term 'constitution' here to refer to the target relation that is under investigation—the relation between, for example, Michaelangelo's *David* and the particular piece of marble that makes it up—then the conclusion of the argument that I shall defend entails that constitution is not identity. The reason that constitution is not identity will emerge in my defense of the argument form (1)-(3). Although a constructive account of constitution will have to await another occasion,[5] my aim here is to vindicate an essentialist argument that shows that constitution is not identity, and to defend the argument in a way that shows why constitution is not identity.

I. A SAMPLE ARGUMENT TO BE DEFENDED

The argument for the conclusion that constitution is not identity can be illustrated by a variation on a justly famous example about a statue from Allan Gibbard (*op. cit.*). Although Gibbard used his example to support contingent identity, I shall use the variation to support a contrary view: constitution without identity. But my overall aim—to defend the argument form (1)-(3)—could be as well served by other illustrative arguments as by the one that I have chosen. So anyone who thinks that concrete things have some of their properties essentially, but that the sample argument about a particular statue is unsound, should select a different illustrative argument. All my arguments in defense of the validity of the statue argument could be deployed, mutatis mutandis, in defense of many other arguments of the form of (1)-(3). I chose an argument concerning a statue as a sample argument largely because of the prominence of statue cases in the literature.

---

across Time," this JOURNAL, LXXX, 4 (April 1983): 201-20; E. J. Lowe, "Instantiation, Identity, and Constitution," *Philosophical Studies*, XLIV (1983): 45-59; Vere Chappell, "Locke on the Ontology of Matter, Living Things and Persons," *Philosophical Studies*, LX (1990): 19-32; David S. Oderberg, "Coincidence under a Sortal," *Philosophical Review*, CV, 2 (April 1996): 145-71.

[4] For an essentialist argument, see Michael Della Rocca, "Essentialists and Essentialism," this JOURNAL, XCIII, 4 (April 1996): 186-202. For anti-essentialist arguments, see the philosophers cited in footnote 2.

[5] This project is underway in my "Unity without Identity: A New Look at Constitution" (manuscript, 1996). As I see it, the relation of constitution is asymmetrical as well as irreflexive.

Here is the variation on Gibbard's example. As a matter of actual fact, the Greek sculptor, Myron, cast a statue of a bronze discus thrower in 450 BCE.[6] Unfortunately, the statue, *Discobolus*, has not survived and is known to us only by Roman marble copies. Now, suppose that Myron created *Discobolus* by first casting two pieces of bronze and then welding them together. That is, suppose that *Discobolus* and the piece of bronze that constituted it came into existence at the same instant when the two smaller pieces were welded. Since *Discobolus* is not extant, suppose that *Discobolus* and the bronze piece that constituted it were destroyed together, at the same instant, a century later. Slightly more fancifully, suppose further that while Myron was deciding how much metal alloy to use to bond the two smaller pieces together, he was pondering the single piece that would result from the two smaller pieces after he bonded them. With his attention so riveted on his welding, he declared: "I hereby dub the piece of bronze that will result from my welding 'Bronze Piece', or 'BP' for short." 'BP'—whose reference is fixed by the definite description, 'the piece of bronze that results from Myron's welding at time t'—rigidly refers to that particular bronze piece.

So *Discobolus* is the statue; BP is the piece of bronze that constitutes the statue. *Discobolus* and BP were, we may assume, wholly coincident throughout their entire histories, and they had the same color, shape, location, and so on. So much Gibbard would accept. But I want to argue against Gibbard and his allies that, nevertheless, mere spatiotemporal coincidence is not enough for identity. Here is an instance of the simple and well-worn argument form to which I alluded earlier:

The statue argument:

(4) *Discobolus* is essentially a statue.
(5) BP is not essentially a statue.
∴ (6) BP ≠ *Discobolus*

Taken at face value, the statue argument is obviously valid. What (4) affirms is this: anything that existed and was not a statue (at all times of its existence) would not be *Discobolus*.[7] If (4) is true, then being a statue is a property that a statue cannot lose without going out of existence—just as being a three-sided figure is a property that a triangle cannot lose without going out of existence. Thus, (4) takes 'statue' as a substance sortal (like 'human being'), as opposed to a phase sortal (like 'child'). (4) entails that, if a statue ceases to be a statue, then it (the thing that was a statue) goes out of existence.

---

[6] H. W. Janson, *History of Art* (Englewood Cliffs, NJ: Prentice-Hall, 1962), pp. 105-06. There are interesting philosophical questions about the status of the marble copies; but it would take us too far afield to consider them here.

[7] More formally: $\forall x \forall t[(x = \textit{Discobolus}) \rightarrow \Box (x \text{ exists at } t \rightarrow x \text{ is a statue at } t)])$.

What (5) affirms is this: it is possible that BP exists and is not a statue.[8] According to (5), BP—that very bronze piece—could have existed without being a statue.[9] For example, for all that (5) claims, BP could have been part of an underground plumbing system in a society without representational art. In that case, BP would have existed and would not have been a statue. Let me explain the claim that BP could have existed in a society without representational art. This claim does not presuppose any particular individuation conditions for bronze pieces; it does not assume that BP would cease to exist on losing a molecule or two; nor does it assume the contrary. The claim is indifferent to whether or not the existence of BP depends on its having a particular maker, or its being formed by welding the two particular pieces together, or its maker's intention to make a piece of bronze. The claim only requires that, however bronze pieces are individuated, BP could have existed without being a statue. So, taking the statue argument at face value, *Discobolus* has a property (being essentially a statue) that BP lacks. In that case, by a familiar form of Leibniz's Law, *Discobolus* ≠ BP. Thus, the statue argument, taken at face value, is valid: (4) and (5) entail (6).

Arguments of the form of the statue argument, however, have been barraged by extremely sophisticated objections. My aim is to shore up the statue argument taken at face value by exposing flaws—in some cases surprisingly simple and fundamental—in the objections to arguments of its form. I shall start with two arguments that aim to show that *Discobolus* is identical to BP. Next, I shall turn to the anti-essentialist charge that the statue argument is question begging in that it misconstrues modal predicates (like '...is essentially a statue').[10] Then, after discussing the idea of contingent identity—the relation that, according to anti-essentialists, holds between *Discobolus* and BP—I shall rebut an essentialist argument that also charges the statue argument with being question begging. At the end, I shall turn back to the question of the truth of the premises, and hence of the soundness, of the statue argument.

---

[8] More formally: $\exists x \exists t[(x = \mathrm{BP})\ \&\ \Diamond\ (x \text{ exists at } t \text{ and } x \text{ is not a statue at t})]$.

[9] This way of putting it suggests that BP is in fact a statue. Although I do hold that BP is (predicatively) a statue derivatively, my argument here does not need that assumption. Anyone who thinks that BP is not a statue at all, but agrees with me that constitution is not identity, could modify the phrase to 'could have existed without being or *constituting* a statue'. Ultimately, in my constructive account in "Unity without Identity," I argue that BP borrows the property of being a statue from *Discobolus*. BP is a statue because, and only because, there is something that BP constitutes that is a statue. But this claim is more controversial than what I need for my arguments for the validity of (4)-(6). What is needed for those arguments is only the claim that the property of being a statue is not one that BP has essentially. Those arguments are indifferent to whether BP has the property contingently or BP lacks the property altogether.

II. RESPONSES TO TWO ARGUMENTS FOR IDENTITY

One way to defeat the statue argument is to present a sound argument for the identity of *Discobolus* and BP. I shall consider two arguments that aim to discredit the statue argument.

Argument 1. "If y is a paradigm F and x is intrinsically exactly like y, then x is an F."[11] *Discobolus* is a paradigm statue; and BP is intrinsically exactly like *Discobolus*; so BP is a statue. Since *Discobolus* and BP are spatially coincident, if *Discobolus* $\neq$ BP, then where *Discobolus* is, there are two coincident statues. But it is intolerable to hold that where *Discobolus* is, there are two coincident statues. So, by modus tollens, *Discobolus* = BP.

Reply. In this argument, the premise that carries the ball is this principle:

(7) If y is a paradigm F and x is intrinsically exactly like y, then x is an F.

The argument is unsound, because (7) is false. No one who ever endorsed (7) could have been thinking about statues. For something is a statue in virtue of its relational properties. But it is obviously false that, if x is an F in virtue of its relational properties, and y is intrinsically exactly like x, then y is an F.[12] Anything defined in terms of relational properties—a planet, a U. S. dollar bill, a passport—provides a counterexample to (7). Specifically, artworks like statues are counterexamples to (7). If we look to the philosophy of art, we find that the competing answers to the question—'In virtue of what is x an artwork?'—concern relational properties. Perhaps something is an artwork in virtue of "the artistic enfranchisement of real objects,"[13] or perhaps in virtue of being an artifact "upon which some society or sub-group of a society has conferred the status of candidate for appreciation,"[14] or perhaps in virtue of being a communication of feeling,[15] or perhaps in virtue of being a certain kind of imitation,[16] or perhaps in virtue of being "caused by a feeling or an emotion on the part of its maker,...which it then expresses,"[17] or perhaps in virtue of something else. Whatever is the correct theory of art, the property or properties in virtue of which something is an artwork are relational.[18] The counterexamples show that (7) is false, and hence cannot be used to show that, if *Discobolus* ≠ BP, then where *Discobolus* is, there are two coincident statues. Hence, argument 1 is unsound.

---

[10] See Gibbard; Robinson; Lewis, "Survival and Identity," in Amelie O. Rorty, ed., *The Identities of Persons* (Berkeley: California UP, 1976), pp. 17-40; and Lewis, "Counterparts of Persons and Their Bodies."

[11] Mark Johnston, "Constitution Is Not Identity," *Mind*, CI (January 1992): 89-105, here pp. 97-98. Johnston set out an argument of this form in order to refute it, and Harold W. Noonan—"Constitution Is Identity," *Mind*, CII (January 1993): 133-46—criticized Johnston and defended the argument. Johnston prefaced his statement of this premise by saying: "Take any sort of thing F, then this principle will be plausible." Although both Johnston and Noonan modify this principle, neither proposes any restrictions on the sort of property that 'F' indicates. The arguments on both sides—both Johnston's and Noonan's—are highly technical and complex and different from the one given here.

[12] Note that the argument in the text also shows that the following principle is also false: if y is a paradigm F and x is intrinsically exactly like y and x does not partly overlap any F, then x is an F. Noonan uses that principle in discussing "the problem of the many" in reply to Johnston (*ibid.*, p. 136).

Argument 2. The second argument for the identity of *Discobolus* and BP is this: *Discobolus* and BP "consist of the very same atoms."[19] If *Discobolus* and BP consist of the same atoms, then, if *Discobolus* and BP are not identical, they differ in kind.[20] If *Discobolus* and BP differ in kind, then "there must be something true of [*Discobolus*], but not of [BP] in virtue of which [they differ in kind]." But there is nothing that is true of one but not of the other in virtue of which they differ in kind. Therefore, *Discobolus* and BP are identical.

---

[13] Arthur C. Danto, "The Artworld," this JOURNAL, LXI, 19 (October 15, 1964): 571-84.

[14] George Dickie, "Defining Art," *American Philosophical Quarterly*, VI (1969): 253-56.

[15] Leo Tolstoy, "Art as the Communication of Feeling: From *What Is Art?*" in George Dickie and Richard J. Sclafani, eds., *Aesthetics: A Critical Anthology* (New York: St. Martin's, 1977), pp. 53-82.

[16] The broad tradition deriving from Plato.

[17] Danto, *The Transfiguration of the Commonplace* (Cambridge: Harvard, 1981), p. 26.

[18] I can think of two possible counterexamples: (1) Benedetto Croce's conception of art as located in the artist's head; what the artist produces is a kind of residue of the true artwork; and (2) Clive Bell's and Roger Fry's conception of art as significant form. I do not believe that either of these can be explicated entirely in nonrelational terms. In any case, neither of these is likely to be the correct theory of art.

I am not here denying that BP also has these relational properties; I am only providing counterexamples to (7). In "Unity without Identity," I formulate a notion of borrowing properties, and argue that BP borrows the property of being a statue from *Discobolus*, and hence is not a distinct statue. The present point, however, concerns the truth value of (7).

[19] The argument comes from Michael B. Burke, "Copper Statues and Pieces of Copper: A Challenge to the Standard Account," *Analysis*, LII (1992): 12-17; the quotation is on p. 14. Burke asks in exasperation: "What, then, could make them different in sort?" He sees only two possible answers: (i) they have different histories, and (ii) they have different persistence conditions. Burke argues that neither of these can ground a difference in sort. His discussion does not consider the possibility that a difference in relational properties required for an F and a G may ground a difference in kind.

Reply. The statue argument outright specifies a property which *Discobolus* has but which BP lacks in virtue of which *Discobolus* and BP differ in kind: *Discobolus* is essentially a statue; BP is not essentially a statue.[21] So simply to assert that there is not such a property would just beg the question against the statue argument. What grounds their difference in kind is this: the properties required for something to be *Discobolus*, as I have just argued, are relational properties; but no relational properties (or at least not the same relational properties) are required for something to be BP. If *x* has certain relational properties essentially, but *y* does not have the same relational properties essentially, then it is not surprising that *x* and *y* differ in kind. So there is no mystery that things consisting of the very same atoms can differ in kind.

In short, the difference in kind between *Discobolus* and BP is determined by the difference in the properties required for the existence of *Discobolus* and the existence of BP. Therefore, the premise that "there is nothing that is true of one but not of the other in virtue of which they differ in kind" is false, and argument 2 is unsound.

III. MODAL PROPERTIES

Since argument 1 and argument 2 each has a false premise, neither gives reason to think that there is anything wrong with the statue argument. But there is a different kind of challenge to the statue argument. This challenge tries to drive a wedge between modal predicates and the properties that they denote. I shall consider two versions of this strategy. The first (argument 3) is based on the claim that concrete things have no modal properties. The second (argument 4) is based on the claim that modal predicates are ambiguous.

Argument 3. Concrete things have no modal properties. If concrete things have no modal properties, then *Discobolus* does not have the property of being essentially a statue. In that case, (1) is false and the statue argument is unsound.

Reply. The weight of this argument is carried by the premise that concrete things have no modal properties. What reason is there to accept such a premise? Gibbard has presented a succinct argument to show that there are no modal properties of concrete things:

---

[20] This argument assumes that two things of the same kind cannot be spatially coincident. For a defense of the assumption, see Oderberg.

[21] On what Burke calls "the standard account," *Discobolus* is not a piece of bronze, nor is BP a statue. That is not my view, as I explain in "Unity without Identity." All that is needed for the present argument, however, is the denial that BP is essentially a statue.

(8) "Modal expressions do not apply to concrete things independently of the way that they are designated."
(9) "A property, if it is to be a property, must apply or not apply to a thing independently of the way that it is designated."
∴ (10) "Expressions constructed with modal operators...simply do not give properties of concrete things" (*op. cit.*, p. 201).

The argument is valid, but, I think, unsound. (8) is subject to counterexamples; for modal expressions include not only predicates like 'is essentially a statue', but also many other kinds of predicates. Suppose that a surgeon removes a bullet from a wounded soldier's shoulder, and later presents the bullet to the injured soldier and declares: "This thing could have killed you." Then it seems true of that particular bullet, independently of the way that it is designated, that it could have killed the soldier. In general, predicates ascribing abilities and powers to concrete things, independently of the way that they are designated, entail that modal expressions apply to concrete things. (For example, Alice can swim the English channel.) Many predicates which are not overtly modal expressions and which apply to concrete things presuppose that modal expressions apply to those concrete things. Predicates that attribute to concrete things dispositions ('is courageous', 'is even-tempered'), attitudes ('is afraid of flying', 'believes that winters are cold in Vermont'), probabilities ('has a probability of .5 of turning up heads'), or causal powers ('is lethal') all apply to concrete things only if modal expressions apply to those things independently of the ways that they are designated. So the truth of statements in which modal expressions apply to concrete things just does not, in general, depend on how those things are designated.

Furthermore, statements containing ineliminable modal expressions that apply to concrete things independently of the ways that they are designated seem to play a role in the sciences. For example, Jupiter could have had one more moon than it does; Mars could have been a site where multicellular life developed. Or suppose that an electron gun in a double-slit experiment is slightly disturbed and fires an electron off-target, so to speak. It is true of that particular electron that it could have hit the target, or that it could have had a slightly different velocity. To say that the truth of such statements depends on how things are designated would be to say that truth in the physical sciences can depend on how things are designated. In that case, anyone who assumes that realism requires truth independently of the way things are designated would face the specter of irrealism in the physical sciences. If this is not what philosophers like Gibbard want to say, then they should deny (8).

The difficulty with both (8) and (9) is that each is formulated as a general principle, without restriction to essential properties. (8)-(10) could be recast in a more restricted version that would avoid the counterexamples. For example, (8)-(10) could be replaced by:

(8′) Modal expressions that purport to attribute essential properties do not apply to concrete things independently of the way that they are designated.

(9′) An essential property, if it is to be an essential property, must apply or not apply to a thing independently of the way that it is designated.

∴ (10′) Modal expressions that purport to attribute essential properties...simply do not give essential properties of concrete things.

The strength of the original (8)-(10) was its generality: it begged no questions against the statue argument. Of course, the downside of that generality were the counterexamples to (8). Now the problem with (8′)-(10′) is the opposite of the problem with (8)-(10). (8′)-(10′) would avoid the counterexamples, but at the cost of begging the question against the statue argument; for a proponent of the statue argument would denounce (8′) right off the bat. So if (8′)-(10′) is to be used without begging the question against the statue argument, (8′) requires independent argument. Whether such an argument will be forthcoming for (8′) which does not beg the question against the statue argument remains to be seen.[22] (Note that a proponent of the statue argument need not deny (9′); for an essentialist may well claim that *Discobolus* is essentially a statue independently of the way that it is designated.)

---

[22] One might suppose that (8′) could be motivated by examples like W.V. Quine's "mathematical cyclist"—see *Word and Object* (Cambridge: MIT, 1960), p. 199. As an argument against essentialism, this case has been thoroughly dissected in the literature and found wanting. For example, see Ruth Barcan Marcus, "Essential Attribution," in *Modalities: Philosophical Essays* (New York: Oxford, 1993), p. 54. Marcus also refers the reader to other writers as well (Terence Parsons, Alvin Plantinga, and Richard Cartwright). I have also heard Max Cresswell and Phillip Bricker discuss Quine's example. Moreover, Cresswell pointed out to me that Quine actually took rejection of essentialism as a premise rather than a conclusion, as evidenced by Quine's attempting to discredit quantified modal logic by claiming that it led to "the metaphysical jungle of Aristotelian essentialism"—"Three Grades of Modal Involvement," in *The Ways of Paradox and Other Essays* (New York: Random House, 1966), p. 174.

The upshot of the discussion of argument 3 is this: its controversial premise—that concrete things have no modal properties—is supported by an argument, (8)-(10), that is unsound. Replacement of the false premise by a true one, as in (8′)-(10′), results in an argument that begs the question against the statue argument. In any case, Gibbard's argument does not establish the premise that there are no modal properties of concrete things. So argument 3 does not refute the statue argument.

Gibbard further takes his argument to show that persistence conditions are not genuine properties of concrete things. He takes the persistence criteria of a thing *x* to specify the conditions under which *x* would continue to exist as a particular kind. He says: "In rare cases, at least one thing will be of two different kinds, with different persistence criteria..." (*op. cit.*, p. 195). So, on Gibbard's view, persistence conditions do not attach to (what he takes to be) the concrete thing, *Discobolus*/BP. Rather, *Discobolus*/BP has one set of persistence conditions qua bronze piece and another set of persistence conditions qua statue. Gibbard's argument for this view is the one that I just refuted: persistence conditions are attributed by modal expressions ("*x* would not continue to exist if..."), and modal expressions do not attribute genuine properties of concrete objects; so, persistence conditions are not genuine properties of concrete objects. Since, as we have seen, arguments of this form are unsound, we may well hold that persistence conditions are properties of concrete things; and, indeed, we should.

Surely, there are conditions for the persistence of *x* per se—as opposed to persistence conditions for *x*-as-an-*F*. Persistence conditions of *x* per se specify the varieties of change that *x* can survive and the kinds of change that would destroy *x*. Here is an argument that *Discobolus* (and hence concrete things) per se have persistence conditions:

(11) If *x* exists at *t* and is not eternal, then *x* can cease to exist (and not just cease to be an *F*).[23]

(12) If *x* can cease to exist (and not just cease to be an *F*), then there are conditions under which *x* would cease to exist (and not just cease to be an *F*), and conditions under which *x* would persist (and not just continue to be an *F*).

(13) If there are conditions under which *x* would cease to exist (and not just cease to be an *F*), and conditions under which *x* would

---

[23] Of course, if being an *F* is an essential property of *x*, then *x*'s ceasing to be an *F* is sufficient for *x*'s ceasing to exist per se.

persist (and not just continue to be an F), then x per se has persistence conditions.

(14) *Discobolus* exists at t and is not eternal (indeed, *Discobolus* existed and then ceased to exist).

∴ (15) *Discobolus* per se has persistence conditions.

We have seen that Gibbard argued (unsuccessfully, in my opinion) against claims like (15). But it is worth pointing out why (15) is significant. (15) cannot be accommodated by the contingent-identity view. On the one hand, it follows from Gibbard's contingent-identity view that:

(16) There are persistence conditions C and C' such that *Discobolus*/BP (qua statue) has C, and *Discobolus*/BP (qua bronze piece) has C', and $C \neq C'$.[14]

On the other hand, if (15) is true, and *Discobolus* = BP, then *Discobolus*/BP cannot have more than one set of persistence conditions. (Otherwise, there would be a circumstance in which, on one of the persistence conditions, x would survive, and on the other of the persistence conditions, x would not survive. But it is impossible for there to be a circumstance in which both x would survive and x would not survive.) So it follows from (15), together with the thesis that *Discobolus* = BP, that:

(17) If there are persistence conditions C and C' such that *Discobolus*/BP has C and *Discobolus*/BP has C', then $C = C'$.

(16) and (17) cannot both be true. So the contingent-identity theory and (15) cannot both be true. Since (15) is entailed by (11)-(14), a contingent-identity theorist would have to show that one of the premises (11)-(14) is false in order to secure the coherence of the contingent-identity view. Since (11)-(14) seem unassailable to me, the coherence of Gibbard's contingent-identity view looks to be in jeopardy.

To sum up my reply to Gibbard's view that concrete things do not have modal properties, I tried to show that Gibbard's argument, (8)-(10), is unsound and hence cannot be used against the statue argument, which attributes modal properties to concrete things. Moreover, I argued that concrete things per se have persistence conditions, where persistence conditions are attributed by modal expressions. Finally, I argued for the significance of the claim that concrete things per se have persistence conditions by showing that that claim is incompatible with the contingent-identity view as developed by Gibbard.

---

[14] Gibbard is explicit on this point. He provides different persistence conditions for statues and pieces of clay, and says: "The persistence criteria that I have given make it clear that often the two [a statue and a piece of clay that constitutes it] are distinct" (*op. cit.*, pp. 190, 188). When a statue and a piece of clay are identical, on this view, the fact that the statue and the piece of clay have different persistence conditions makes their identity contingent.

As we have just seen, argument 3 attacks the statue argument by exploiting the difference between modal predicates and modal properties. Other arguments exploit the same difference, by claiming that modal predicates are ambiguous.[25] For example, some claim that de re modal predicates are "predicates whose reference is affected by the subject term to which they are attached."[26] Harold Noonan, who calls such predicates *Abelardian predicates*, says that the view of modal predicates to which "the defender of contingent identity is committed is that modal predicates are Abelardian predicates whose reference is determined by a component of the sense of the subject expression to which they are attached" (*ibid.*, pp. 189-90).[27] On this view, modal predicates of the form 'is essentially a statue' do not attribute a single property in all linguistic contexts. Now let me give a Noonan-style argument against the statue argument.

Argument 4. The property denoted by 'is essentially a statue' in (1) is not the same property as the property denoted by 'is essentially a statue' in (2). So there is not a single property that is attributed to *Discobolus* and not attributed to BP. If there is not a single property that *Discobolus* has but BP lacks, then (1) and (2) do not entail (3)—and the statue argument is invalid.

Reply. The first premise depends on the claim that the property denoted by '...is essentially a statue' depends on the meaning of the subject term to which it is attached. I believe that this key claim is false. My argument here is extremely simple: expressions denoting persistence conditions have the same status as expressions denoting essential properties, with respect to dependence on the meanings of subject terms to which they are attached. In that case:

(18) (The meaning of a predicate of the form 'is essentially F' depends on the meaning of the subject term to which it is attached) if and only if (the meaning of a predicate expressing persistence conditions depends on the meaning of the subject term to which it is attached).

---

[25] Lewis is perhaps the most prominent proponent of the view that de re modal predicates are ambiguous. See his "Survival and Identity" and his "Counterparts of Persons and Their Bodies."

[26] Noonan, "Indeterminate Identity, Contingent Identity and Abelardian Predicates," *The Philosophical Quarterly*, XLI, 163 (1991): 183-93, here p. 188

[27] Noonan points out that Lewis's counterpart theory (as revised to admit a variety of counterpart relations) provides one way of putting flesh on the bones of the idea that modal predicates are Abelardian predicates, but "stress[es] that it is only to the skeletal idea that modal predicates are Abelardian and not to its counterpart-theoretic interpretation that the defender of contingent identity is committed" (*ibid.*, p. 190).

I have just argued, however, that concrete things per se have persistence conditions. (For example, it is not just that *Discobolus* has one set of persistence conditions relative to being a statue named '*Discobolus*', and another set of persistence conditions relative to being a piece of bronze named 'BP'.) In that case:

(19) It is false that the meaning of a predicate expressing persistence conditions depends on the meaning of the subject term to which it is attached.

From (18) and (19), it follows that:

(20) It is false that the meaning of a predicate of the form 'is essentially F' depends on the meaning of the subject term to which it is attached.

Thus, I believe that the statue argument stands against all four of the counterarguments I have canvased. Now I shall turn to a central motivation for counterarguments.

IV. CONTINGENT IDENTITY

One motivation for holding that BP is identical with *Discobolus* is a strong intuition shared by many philosophers. It is the intuition that some things that are in fact identical might not have been identical. Granted, the intuition goes, BP and *Discobolus* might have been distinct; but in fact they are not.[28] They are contingently identical:[29]

(21) x is contingently identical to $y =_{df} (x = y)$ & $\Diamond(x$ exists & y exists & $x \neq y)$

Now, traditionally, identity has been understood as a necessary relation for which the following *thesis of the necessity of identity* holds with full generality:

(22) $x = y \rightarrow \Box\,(x = y)$

Those who endorse the thesis of the necessity of identity would deny that anything satisfies the above definition of 'contingent identity'. Michael Jubien[30] put it well:

> Undoubtedly there are some relations that behave in some respects like identity but which do not hold of necessity. I believe it is a fundamental

---

[28] Such a contingent-identity theorist may bolster his intuition with a theory of modal predication of the kind already considered.

[29] Gibbard, p. 187.

[30] *Ontology, Modality and the Fallacy of Reference* (New York: Cambridge, 1993). Agreement with Jubien on the nature of identity, of course, does not entail agreement about what in fact exists.

> error to think that one of these relations actually is the identity relation, that is, to think that identity actually doesn't hold of necessity. I also believe it is seriously misleading to agree that none of these relations is the identity relation, but nevertheless to dub one of them 'contingent identity' (*ibid.*, p. 39).

My own position is that identity is necessary, and that no objects satisfy the definition of 'contingent identity' given in (21); and, further, that constitution, our target relation, is not identity of any sort. Constitution is rather one of those relations "that behave in some respects like identity, but do not hold of necessity."[31]

Here we have a clash of the deepest of intuitions—intuitions that may seem to be (but are not quite) beyond the reach of argument. Does it even make sense to say that *a* is identical to *b*, but that *a* might not have been identical to *b*? Although reasons are difficult to come by at this level of intuition, I believe that—even if the contingent-identity theory exposed some flaw in the argument for (15)—there would remain good reason to prefer constitution-without-identity; for the contingent-identity view of constitution has theoretically unsatisfying consequences that a nonidentity view of constitution lacks. One of these infelicitous consequences is metaphysical; the other is epistemological.

First, consider a metaphysically untoward consequence of the contingent-identity view: the contingent-identity view affords no unified account of the relation between persons and bodies, nor of the relation between statues and pieces of bronze, and so on. The contingent-identity theorist holds, roughly, that *a* and *b* are contingently identical only if *a* and *b* share all their categorical properties, where categorical properties are those which do not depend on how things are in other worlds.[32] For example, if *a* and *b* are contingently identical, then *a* and *b* start to exist at the same time and cease to exist at the same time. Gibbard suggests that a person may cease to exist when she dies. If Smith's body continues to exist (as a corpse), then—everyone would agree—Smith, the person, is not contingently identical to her body. But if the body is destroyed at death, then "there is no purely logical reason against saying the following: the person is this case is identical with his body, but had he died a normal death, he would have been distinct from his body" (*op. cit.*, p. 213).[8] On Gibbard's construal of contingent identity, then, whether or not a person is contingently identical to her body may depend on a fluke at her death: if she leaves a corpse, she is not contingently identical to her body; if she does not leave a corpse, then there is no logical reason to deny that she is identical to her body. Therefore, this contingent-identity view does not assert a uniform relation between all persons and their bodies—nor between statues and pieces of bronze that make them up, nor between any of the other things that, on my view, are related by constitution. Instead, the contingent-identity view of constitution interjects needless bifurcations into our conception of the world.

---

[31] Partly "for shock value," Yablo uses the term 'contingent identity' for the relation that I am calling 'constitution' (*op. cit.*, p. 303). But he disavows the assumption that "contingently identical things were (at least) properly identical, only—and this was their distinction—not necessarily so." That is, Yablo is not using 'contingent identity' in the sense of (21). I think that it is less misleading to avoid the term 'contingent identity' altogether.

[32] Yablo brings to light difficulties in formulating a clear, noncircular statement of a thesis of contingent identity (*op. cit.*).

Surely, contrary to the contingent-identity view of constitution, all human persons have the same relation to their bodies, whatever it is. If I am identical to my body, so are you to yours; and if I am not identical to my body, you are not identical to yours either. Similarly, there should be unified accounts of the relation between statues and pieces of marble (or gold or whatever), and of the relation between flags and pieces of cloth, and so on, as well. Constitution-without-identity is superior to constitution-as-identity in that it provides a unified view of the relation between persons and bodies, statues and pieces of bronze, and so on.

We can see this point from another angle. If we pretheoretically understand constitution to be the relation between, say, statues and pieces of bronze that make them up, then, on the contingent-identity view, in most cases, constitution is not identity. Gibbard is explicit on this point:

> In a typical case, a piece of clay is brought into existence by breaking it off from a bigger piece of clay. It then gets shaped into the form of an elephant. With the finishing touches, a statue of an elephant comes into being. The statue and the piece of clay therefore have different properties: the times they start to exist are different, and whereas the statue has the property of being elephant-shaped as long as it exists, the piece of clay does not. Since one has properties the other lacks, the two are not identical (*op. cit.*, p. 190).

Now, in the case just described by Gibbard, a piece of clay (call it 'Clay') constitutes a statue (call it 'Elephant'), without being identical to it. So, to describe the relation between Clay and Elephant, even proponents of contingent identity would need recourse to some notion in addition to contingent identity anyway.

---

[8] Gibbard continues: "If there are any reasons against such a view, they must be non-logical reasons." So, he is not actually committed to the view that whether or not one is identical to one's body depends on a fluke at death. Still, the possibility that he envisages is, I believe, theoretically unacceptable.

So the contingent-identity theorist is in the odd position of saying that the relation between Elephant and Clay is not the same as the relation between *Discobolus* and BP (as I have imagined it). But, surely, it is at least a desideratum to have a single account of the Elephant/Clay relation and the *Discobolus*/BP relation. The notion of constitution-without-identity, in contrast to the notion of contingent identity, allows that desideratum to be satisfied. Therefore, the notion of contingent identity cannot do all the work that the notion of constitution does.

The second theoretically untoward consequence of the contingent-identity view of constitution is epistemological: if the contingent-identity view were correct, we would typically not be justified to assert of *a* and *b* that they are contingently identical while they/it exist(s). This is so, because in order for the identity of *a* and *b* to be contingent, it must be possible that they have different properties; but in order for *a* and *b* to be identical at all, *a* and *b* must actually have all their properties in common—including ceasing to exist at the same time. But before the demise of *a* or *b*, we do not know whether they will differ in the future (and hence not be identical). For example, I have no idea how I shall die; I do not know whether I shall leave a corpse or not. Hence, on the contingent-identity view, I lack justification for supposing that I am or that I am not contingently identical to my body. Only after I die is a contingent-identity theorist in a position to declare whether I am contingently identical to my body; presumably, I shall never know. On a nonidentity view of constitution, if *x* constitutes *y* (now), we can be justified in asserting now that *x* constitutes *y*; but on the contingent-identity view, if *x* is contingently identical to *y* (now), we are typically not justified in asserting now that *x* is contingently identical to *y*.

Thus, there are both metaphysical and epistemological reasons to prefer a nonidentity view of constitution to a contingent-identity view. Moreover, the purposes to be served by the idea of contingent identity are as well served without recourse to contingent identity: for example, statements like 'The inventor of bifocals was the first U.S. Postmaster General' are true and contingent, but they do not express any contingent identity in the sense defined by (21).[34] In short, not only are the arguments against the statue argument resistible, but also the motivation for those counterarguments can be undercut as well.

---

[34] Using Bertrand Russell's theory of descriptions, the terms that seem to refer to individuals (for example, 'the inventor of bifocals') disappear in favor of variables and predicates.

V. AN ESSENTIALIST OBJECTION TO THE STATUE ARGUMENT

Even if we reject the contingent-identity view of constitution, there remains another threat to the statue argument: constitution as necessary identity. An essentialist who endorses the thesis of the necessity of identity may yet reject the statue argument in favor of this argument:

(23) *Discobolus* is essentially a statue.
(24) *Discobolus* = BP
∴ (25) BP is essentially a statue.

Is there any reason to prefer the statue argument over (23)-(25)? Does each just beg the question against the other? I think not. I think that the argument that BP is not essentially a statue is stronger than the argument that *Discobolus* = BP. But the argument that BP is not essentially a statue depends on modal intuitions: BP could have existed without being a statue. For example, BP could have existed as a piece in an underground plumbing system in a world without art. Hence, BP is not essentially a statue.

Recently, arguments of the form of the statue argument—supported by modal intutions like those just expressed—have come under fire. Michael Della Rocca has argued that, under certain natural assumptions, all arguments of the form of (1)-(3) are question begging (*op. cit.*).[23] This is a strong conclusion, indeed, and, I shall argue, a mistaken one. If certain arguments of the form (1)-(3) are sound—and I shall try to show that they are—then it is clear that constitution is not identity.

Della Rocca's argument is designed to show that arguments of the form of the statue argument fail even from an essentialist point of view. The essentialist point of view here is Kripkean essentialism. Say that a Kripkean essentialist is one who (i) thinks that there are certain properties and certain things such that those things cannot exist without exemplifying those properties, and (ii) accepts Saul Kripke's distinction between the meaning of a term and the fixing of its reference. Consider an argument that has the same face-value form as the statue argument:

---

[23] I modify Della Rocca's argument to make it applicable to the statue argument.

The motion argument:

(26) Molecular motion is essentially molecular motion.
(27) Heat is not essentially molecular motion.
∴ (28) Heat is not identical with molecular motion.

Kripke has a well-known argument to show that (28) is not established by the premises of the motion argument, and Della Rocca argues that there is no good reason to refrain from using similar considerations to show that (6) is not established by the premises of the statue argument. Della Rocca focuses on Kripke's reconstrual of the second premise of the motion argument. (27) does not properly express our intuition about heat; rather, if we think harder (and distinguish between reference fixing and meaning), we shall see that the intuition should not be expressed as (27). Instead, assuming that we actually fix the reference of 'heat' by 'the cause of sensation *S*', where sensation *S* is a heat sensation, our intuition about heat should be expressed as (29):

(29) There is a possible situation in which something which is not heat but which produces in us sensation *S* fails to be molecular motion (*op. cit.*, p. 192).

But the argument resulting from substituting (29) for (27) in the motion argument is invalid.

Now, Della Rocca notes, there is an analogous reconstrual of (5) available for the statue argument. Assuming that the reference of 'BP' is fixed by the definite description, 'the one and only piece of bronze that results from Myron's welding at time *t*', the parallel reconstrual of (5) is (30):

(30) There is a possible situation in which an object that is not BP, but has the property of being the one and only piece of bronze that results from Myron's welding at time *t*, is not a statue.

Now "for an essentialist," says Della Rocca, "our intuition in the matter of heat can be expressed in two ways—one way [(29)] compatible with the identity of heat and molecular motion and one way [(27)] not compatible with this identity. From this it follows that an essentialist would not allow one to argue simply from our modal intuition in this matter to a conclusion of nonidentity" (*op. cit.*, p. 196).[illegible] Simi-

---

[illegible] Della Rocca supposes that there is a single intuition that may be expressed in either of two ways—(27) or (29); he says that "the intuition of contingency originally expressed by [(27)] could also be captured by [(29)]" (*op. cit.*, p. 198). But surely, since (27) and (29) manifestly "say" different things, they are not just two ways of expressing a single intuition. However intuitions are individuated, since (27) is false and (29) is true, it would be more plausible to say that there are two larly, on Della Roccan reasoning, our intuition about bronze pieces can be expressed in two ways—as (5) or as (30). So, given the availability of (30), we need a principled reason, says Della Rocca, to consider (5) as the correct interpretation of the intuition underlying the second premise. Therefore, says Della Rocca, the statue argument is question begging—unless there is a reason to reject (30) as a reconstrual of (5). The only available reason, he thinks, would involve prior assumption of the nonidentity of BP and *Discobolus*. Obviously, we cannot assume the nonidentity of BP and *Discobolus* in an argument for their nonidentity (*op. cit.*, p. 196). If this is so, then the statue argument is question begging, period. Here is a more compressed version of Della Rocca's argument:

Argument 5. When (27)—which, on Kripkean essentialism, is false—is replaced by its reconstrual (29) (which is true), the revised motion argument is invalid. There is no good reason to refrain from replacing (5) by its reconstrual (30) in the statue argument in the same way. If there is no good reason to refrain from replacing (5) in the statue argument in the same way that (27) was replaced in the motion argument, it is question begging to take the statue argument at face value as a valid argument. Thus, it is question begging to take the statue argument at face value as a valid argument.

Reply. Argument 5, too, is unsound; for Della Rocca's second premise—that there is no good reason to refrain from replacing (5) by its reconstrual (30) in the statue argument—is not in the slightest justified by consideration of the flaw in the motion argument. What is wrong with the motion argument is that, given Kripkean essentialism, one of its premises—(27)—is false, independently of any considerations about reconstrual; but, as I shall argue momentarily, Kripkean essentialism provides good reason to hold that the parallel premise in the statue argument—(5)—is *true*. Therefore, there is very good reason to refrain from replacing (5) in the statue argument.

What motivates reconstrual and replacement of (27) is that (27) is false; yet we have an intuition that (27) seems to express. The point of the reconstrual of (27) is to show that one's mistake in holding (27) is understandable.[37] Reconstrual of (27) answered a question that we had about the motion argument: "Why does (27), which is

---

intuitions, one correct and one incorrect, and one (correctly) expressed by (29) and one (correctly) expressed by (27). But my argument does not depend on contesting Della Rocca's way of counting intuitions.

[37] I am simply saying that a Kripkean essentialist has a way to avoid Della Rocca's charge; I am not claiming that Kripke himself would approve.

false, *seem* to be true?" But without an independent argument to show that (5) is false, the parallel question regarding the statue argument just does not arise. In this case, the second premise of argument 5—there is no good reason to refrain from replacing (5) by its reconstrual (30) in the statue argument in the same way that (27) was replaced by (29)—is unjustified. For in the absence of an independent argument against (5)—an argument similar to the Kripkean argument against (27)—there is no motivation to reconstrue (5). Thus, without the (unargued for) assumption that (5) is false, then argument 5 is unsound, and the statue argument, taken at face value, still stands.

Not only is the motivation for replacing (5) missing, but also the same essentialist tack that showed that (27) is false can be deployed to show that (5) is true. That is, the essentialist who argued that, given her essentialism, (27) is false and hence subject to replacement by a reconstrual, also has resources to show that, given her essentialism, (5) is true and hence not subject to replacement by a reconstrual. Consider this form of essentialism: for every concrete thing, there is a kind of which the thing fundamentally is a member. No concrete thing is fundamentally a member of more than one kind. As Aristotle might say, the kind that provides the answer to the question—'What is *x*?'[illegible]—is the kind of which *x* is fundamentally a member. The essential kind-properties of a thing are the essential properties that all members of its fundamental kind share; if a thing has any other essential properties, then either they are entailed by its essential kind-properties, or they are not properties essential to all members of any kind.[illegible]

Such an essentialist then may say: BP's fundamental kind is the kind bronze. So BP's essential kind-properties are the essential properties of pieces of bronze and properties that are entailed by essential properties of pieces of bronze. BP has no other kind-properties essentially. Being a statue is a kind-property that is not entailed by the essential properties shared by all pieces of bronze. Otherwise,

---

[illegible] See also Wiggins's thesis of sortal dependency (*op. cit.*, ch. 2). According to Wiggins, the answer to this question has "both a sortal component (*What is x? It is an F thing*) and a deictic or particularizing component (*Which F thing is x? It is that F; or It is the F which is φ*)" (*op. cit.*, p. 115).

[illegible] For example, *G* is entailed by *F* if and only if necessarily all *F*s are *G*. (If Tiny is of the fundamental kind elephant, and Tiny has the property of being an animal essentially, then the property of being an animal is entailed by the property of being an elephant.) I remain neutral here on Kripke's doctrine of the essentiality of origin.

every piece of bronze would be a statue. Therefore, BP does not have the property of being a statue essentially. That is, (5) is true.

Note that the argument that (5) is true depends only on general considerations about properties: for example, the property of being a statue is a kind-property that is not an essential property of pieces of bronze; nor is it entailed by the essential properties shared by all pieces of bronze. No particular (non)identity claims of the form $x \neq y$ are assumed. Now, according to Della Rocca, "we can know that x and y differ with regard to [a modal] property only if we already know that x and y are not identical" (*op. cit.*, p. 202). On the version of essentialism I just sketched, however, it is on the basis of *general knowledge* of statues and bronze pieces that we know that a particular statue and a particular bronze piece differ in regard to a modal property. Antecedent knowledge of the identity or nonidentity of x and y does not come into it. So the argument for the truth of (5)—the second premise in the statue argument—begs no questions.

I think that Della Rocca made a subtle two-fold mistake: first, by focusing on the invalidity of the argument in which (27) is replaced by (29), he underestimated the real problem with the motion argument—namely, that it is unsound because (27) is false; second, he took the mere availability of a reconstrual of (5) to motivate replacement of (5) in the statue argument. But the reason to replace (27) in the motion argument was not the mere availability of a reconstrual, but rather the fact that (27) was false. For Della Rocca's analogy to succeed, he must show that (5) is likewise false and the statue argument thus is unsound. Not only does Della Rocca offer no argument for the falsity of (5), but, on the contrary, I have shown that an essentialist has an argument for the truth of (5).

In short, given essentialism, (27) is false, and the motion argument is straightforwardly unsound, independently of any reconstrual of (27); but also given essentialism, (5) is true, and there is no parallel basis for claiming the statue argument to be unsound and no motivation to replace (5) with the admittedly available reconstrual of (5). Hence no question is begged by accepting the statue argument at face value. Therefore, I do not believe that Della Rocca has made good on his claim that arguments of the form of the statue argument are question begging or otherwise unsound.

VI. CONCLUSION

We have surveyed challenges to the statue argument from both the anti-essentialist "left" and the essentialist "right." Both the anti-essentialist and essentialist charge, in different ways, that arguments like the statue argument containing modal predicates, are question begging. I hope to have shown that the statue argument survives the challenges. If I have, then I think that the statue argument should be taken at face value as valid. If its premises are true—or if the premises of any other argument of the form (1)-(3), where *y* constitutes *x*, are true—then constitution is not identity, whether identity is construed as a contingent or a necessary relation.

But is the statue argument sound? Since I doubt that there are conclusive non-question-begging arguments for or against its premises, I would settle for a stalemate. Nevertheless, let me press the case for soundness. I have just argued for (5) on essentialist grounds. Indeed, I believe that the only reason to doubt (5) would depend on prior rejection of the conclusion (6). But one can hardly impugn an argument by using the denial of its conclusion to discredit a premise. Here is another angle. Suppose that in anticipation of welding the two pieces of bronze together to form BP, Myron had said: "I don't yet know whether BP will constitute a statue or not. If my welding is unsuccessful, then BP will be just another piece of bronze to be cast aside." If Myron's utterance, taken at face value, is true, then BP could exist without being a statue. In that case, again, (5) is true.

Although (4) is more controversial, I do have a thought experiment in support of (4). If (4) is false, then not only could *Discobolus* exist without being a statue, but also, presumably, all the other artworks that do exist could exist without being artworks. That is, if (4) is false, there is another possible world that contains every individual that actually exists, but not a single artwork. This consequence of denying (4) seems to me unacceptable. Although I realize that I am wielding only an intuition pump (in Daniel Dennett's memorable phrase), it is clear to me that any world without artworks is missing some of the individuals that populate the actual world. Taking ontology to concern what individuals exist, a world without art would be ontologically impoverished compared to our world. Anyone who agrees will accept (4).

Now suppose, as I have urged, that some argument—the statue argument or some other argument—of the form of (1)-(3), where *x* bears constitution relations to *y*, is sound. In that case, constitution is not identity. If the relation between a thing and what constitutes it is not identity, would the relation between *Discobolus* and BP be just an unexplained fact? No. For, as we have seen, *Discobolus* has relational properties essentially that BP does not have essentially. If there is some relational property in virtue of which *y* is the thing that it is, but not in virtue of which *x* is the thing that it is, then *x* and *y* are not identical. So the nonidentity of *Discobolus* and BP is understandable.

There is a longstanding tradition of assuming that all essential properties of a thing are nonrelational. (The prejudice against relational properties is abetted by equivocal use of the term 'intrinsic' to mean both 'nonrelational' and 'part of the nature of its bearer'.) Even if it turns out that fundamental physical particles have only nonrelational properties essentially, it would not follow that "higher-level" objects have only nonrelational properties essentially.[*] Not only would refusal to countenance relational properties as essential be arbitrary, but also it would rule out a priori the very cases of interest—namely, objects defined by their relational and/or intentional properties. In any case, simply to assume that things do not have intentional or relational properties essentially would be no refutation of the statue argument.

In conclusion, I hope to have shown that the idea of constitution-without-identity is preferable to constitution-as-identity. It remains now to say just what constitution is. Although I must save that inquiry for another occasion, let me say in advance why constitution is so important. If I am right about constitution-without-identity, then not only do subatomic particles exist, but also so do all manner of things of other kinds—things whose properties are not determined by the properties of the subatomic particles that constitute them. In the natural world there are planets, kidneys, persons, landscape paintings, carburetors, cathedrals. None of these things is what it is (the thing that it is) in virtue of its intrinsic properties. Each of them is constituted by another thing (ultimately by an aggregation of subatomic particles) with which it is not identical. So it is well worthwhile to try to discover how constitution actually works.

LYNNE RUDDER BAKER

University of Massachusetts/Amherst

---

[*] This is the lesson of "externalism" in the philosophy of psychology. See Tyler Burge, "Individualism and Psychology," *Philosophical Review*, xcv, 1 (January 1986): 3-46; Jerry Fodor, *The Elm and the Expert* (Cambridge: MIT, 1994); and my *Explaining Attitudes: A Practical Approach to the Mind* (New York: Cambridge, 1995), ch. 2.

HOW TO KEEP THE 'PHYSICAL' IN PHYSICALISM*

Physicalism is roughly the thesis (1) that every entity is either itself a physical entity or is exhaustively composed, ultimately, of physical entities, and (2) that every property is either itself a physical property or is realized, ultimately, by physical properties.[1] Never mind whether exhaustive composition and realization are the best notions for a physicalist to employ in explicating the idea that everything is physical—better, for instance, than the notions of token identity and supervenience;[2] and never mind also how these notions are to be precisely understood. The problem I wish to discuss is what physicalists can and should mean by 'physical' in their formulations of physicalism, and this problem arises on any reasonable view as to the answers to these questions.

A physicalist's definition of 'physical' can perfectly well be stipulative—the problem is not one of conceptual analysis—but to be plausible it must at least meet the following two conditions: when plugged into an otherwise satisfactory formulation of physicalism, it must yield a thesis (1) that is not obviously false and (2) that possesses content determinable by us. (A physicalism whose content was not determinable by us would presumably be impossible for us to support empirically, and might, for all we know, not even exclude from existence the sort of paradigmatically nonphysical items—for example, souls, entelechies, ghosts—which physicalists have traditionally refused to countenance.) There is a dilemma, however, apparently owed to Carl G. Hempel,[3] that is sometimes thought to show that no physicalist definition of 'physical' can

---

* I am grateful for helpful comments on earlier drafts to Geoffrey Hellman, William Lycan, Peter Markle, and Paul Weirich.

[1] See, for example, my "Being a Physicalist: How and (More Importantly) Why," *Philosophical Studies*, LXXIV (1994): 221-41.

[2] See, for example, John F. Post, *The Faces of Existence* (Ithaca: Cornell, 1987).

[3] "Comments on Goodman's *Ways of Worldmaking*," *Synthese*, XLV (1980): 193-99, especially pp. 194-95. See also his "Reduction: Ontological and Linguistic Facets," in Sidney Morgenbesser, Patrick Suppes, and Morton White, eds., *Philosophy, Science, and Method: Essays In Honor of Ernest Nagel* (New York: St. Martin's, 1969), especially pp. 180-83. Compare J. J. C. Smart, "The Content of Physicalism," *Philosophical Quarterly*, XXVIII (1978): 339-41; and Noam Chomsky, *Language and Mind* (New York: Harcourt Brace Jovanovich, 1972), p. 98.

0022-362X/97/9412/622-37

meet these two conditions simultaneously.[4] It is nicely expressed by Geoffrey Hellman:[5]

> ...current physics is surely incomplete (even in its ontology) as well as inaccurate (in its laws). This poses a dilemma: either physicalist principles are based on current physics, in which case there is every reason to think they are false; or else they are not, in which case it is, at best, difficult to interpret them, since they are based on a "physics" that does not exist—yet we lack any general criterion of "physical object, property, or law" framed independently of existing physical theory (*ibid.*, p. 609).

So if "physical" entities and properties are those mentioned as such in the laws and theories of *current* physics, then physicalism is very probably false; but if they are those mentioned as such in the laws and theories of *completed* physics, then, since we have no idea what completed physics will look like, the resulting formulation of physicalism will lack content determinable by us. I shall argue here that the first horn of this Hempelian dilemma is blunt, and that therefore it remains open to a physicalist to understand by "physical" entities and properties those mentioned as such in the laws and theories of *current* physics.[6]

I

Here is what is supposed to be wrong with defining 'physical' in terms of current physics. Past theories in physics, when judged from the standpoint of current physics, have usually turned out to be both false and incomplete; it is therefore very likely (though not, of course, absolutely certain) that current physics is both false and incomplete.[7] But if so, and if physicalism formulated in terms of current physics assumes the truth and completeness of current physics, then it is very likely (though not, of course, absolutely certain) that physicalism is false, too[4]—which requires one to cease to be a physicalist.

---

[4] See, for example, T. Crane and D. H. Mellor, "There is No Question of Physicalism," *Mind*, XC (1990): 185-206, especially p. 188; and Crane, "Why Indeed? Papineau on Supervenience," *Analysis*, LI (1991): 32-37, especially p. 34. See also Bas C. van Fraassen, "Science, Materialism, and False Consciousness," in Jonathan L. Kvanvig, ed., *Warrant In Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge* (Lanham, MD: Rowman and Littlefield, 1996), especially pp. 163-70 and 173-74.

[5] "Determination and Logical Truth," this JOURNAL, LXXXII, 11 (November 1985): 607-16, p. 609.

[6] For a critical survey of other responses to the dilemma, plus a positive proposal very different from mine, see Jeffrey Poland, *Physicalism: The Philosophical Foundations* (New York: Oxford, 1994), ch. 3.

[7] It is not entirely clear what an advocate of Hempel's dilemma should or even could mean by 'incomplete' here; but I assume the intuitive idea is that current physics is incomplete if and only if it fails to mention some entity or property which (a) exists and which (b) would, if discovered, unhesitatingly be classified as physical—such a thing as a new particle with mass, charge, and spin. A difficulty with such an account, of course, is that it *presupposes*, contrary to the conclusion of the dilemma, that a viable conception of 'physical' is now available to us, but an advocate of the dilemma can no doubt get around this problem.

My reply to this argument is to challenge its final step, that is, the inference that a physicalist should abandon physicalism just because physicalism is very likely false. The argument assumes that a physicalist is someone who must assign a high, or even very high, probability to the thesis of physicalism; wherefore it is unreasonable for a physicalist to define 'physical' in terms of current physics if doing so implies that physicalism is improbable. But I deny this assumption, claiming that a physicalist need *not* assign a high probability to physicalism, and can therefore comfortably live with the result that physicalism has a very low probability. There are different ways of developing such a reply to the argument, however, ways that vary in their positive account of what attitude to physicalism a physicalist should take. For instance, one could say that a physicalist is someone who merely treats the thesis of physicalism as some sort of regulative ideal for science, a role it could play while being literally false; and, indeed, Hellman seems to say something just like this in his own response to Hempel's dilemma (*op. cit.*, p. 610). Alternatively, one could say that a physicalist is someone who holds that physicalism, while literally false, is nevertheless closer to the truth, a better approximation to the truth, than its rivals. But both these suggestions have drawbacks. The first requires us to abandon the intuition that a physicalist is someone who takes some sort of attitude toward the *truth* of physicalism. The second can only be as good as the account of verisimilitude or approximation to the truth on which it relies, and these notions are notoriously hard to explicate satisfactorily. By contrast, the account I shall give here respects the intuition that a physicalist is someone who takes some sort of attitude toward the truth of physicalism; and it has no dependence on the concept of verisimilitude.

---

[4] Of course, the past history of physical theorizing is not the only evidence relevant to current physics' likelihood of being true, since there is also independent (that is, nonhistorical, observational) evidence for the claim that current physics is true to take into account; and when this other evidence is taken into account, it is far from obvious that current physics will emerge on balance as very unlikely (see Michael Levin, "On Theory-Change and Meaning-Change," *Philosophy of Science*, XLVI (1979): 407-24, especially pp. 420-21). But I waive Levin's point in what follows. One reason for doing so is that there seems to be no independent evidence for the claim that physics is *complete* which could outweigh the historical evidence for the claim that it is incomplete; so Levin's point could not be the whole of a response to the first horn of Hempel's dilemma.

My development of a different basis for the claim that a physicalist need not regard physicalism as more probable than not will involve giving an account, not of the thesis of physicalism, but of what it is to be a physicalist. Here is my argument in outline.

(P1) To be a physicalist is to take the same attitude—whatever that attitude is—toward the hypothesis of physicalism which those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses.

(P2) The attitude that those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses is identical with an attitude (to be defined later) that I shall call the *SR attitude*.

(C1) Therefore, to be a physicalist is to take the SR attitude toward physicalism.

(P3) But to take the SR attitude toward a hypothesis does not require regarding it as likely to be true (let alone very likely to be true).

(C2) Therefore, to be a physicalist does not require regarding physicalism as likely to be true (let alone very likely to be true).

(P1) is very plausible, once physicalism is viewed as no more and no less than a scientific hypothesis, albeit one with unusual features. Surely, it can be so viewed. It does not, admittedly, assert the holding of any law, nor does it belong to any *one* science. But it does make a claim concerning the constitution/realization of the world which, like the claim that genes are made of DNA (though obviously on a larger scale), is entirely contingent and which is surely scientific rather than, say, commonsensical or religious. As I shall show at the end of my final section, physicalism can be exhibited as being explanatory in just the way we expect scientific hypotheses to be. At any rate, if physicalism is not a scientific hypothesis, then it resembles one in every respect relevant to current purposes. (P3) will turn out to be a triviality, once the nature of the SR attitude is made clear, as will shortly be done. (P2), however, which urges the identification of attitudes not obviously identical, is far from trivial and will require much elucidation and defense.

II

The elucidation and defense of both (P2) and (P3) must begin with an explanation of what I mean by the 'SR attitude' toward a hypothesis, however. Here is a *stipulative* definition:

(SR) To take the SR attitude toward a hypothesis is (1) to regard the hypothesis as true or false in virtue of the way the mind-independent world is, and (2) to assign the hypothesis a higher probability than that of its *relevant rivals*.

But what are the relevant rivals to a hypothesis? Here is another stipulative definition:

> (RR) Hypothesis H1 is a relevant rival to H2 if and only if (a) H1 is sensibly intended to achieve a significant number of H2's theoretical goals; (b) the hypotheses, H1 and H2, fail to supervene on one another; and (c) H1 has actually been formulated.

Clauses (a) through (c) require some unpacking and motivation. Take clause (a) first. The theoretical goals of a hypothesis will include such things as the satisfactory explanation of certain phenomena and the solution of certain problems. The reference to a "significant number" of such shared goals is supposed to do (rough) justice to the fact that advocates of rival hypotheses almost never completely agree what theoretical goals their respective hypotheses can reasonably be expected to achieve; insisting that *all* theoretical goals be shared for two hypotheses to count as rivals would therefore leave almost no pairs of hypotheses as rivals. The reference to goals that are "sensibly" intended for a hypothesis to achieve is supposed to ensure that the hypothesis that, say, the moon is made of green cheese does not qualify as a relevant rival to Darwinism just because some lunatic thinks it can account for the origin of species. Let us now turn to clause (b). It is included so that hypotheses at different levels of explanation (for example, folk psychology and scientific psychology), whose theoretical goals arguably overlap considerably, are not mistakenly classified as rivals; presumably folk psychology does supervene upon scientific psychology. It also serves to exclude hypotheses that are merely notational variants of one another from counting as relevant rivals. Finally, clause (c). I have no full account of what it is for a hypothesis to be formulated, but two points are crucial. First, to count as formulated, a hypothesis need not have been formulated in any great detail, but we must have been told something about its basic principles. So, for instance, I would count creationism as having been formulated, on the grounds that we are told *something* about the basic mechanism it hypothesizes to account for life, even though creationists are notoriously stingy with suggestions as to the details of God's plans for flora and fauna. Secondly, we must distinguish formulating a hypothesis from referring to it. The expression 'the set of laws accepted by people at Harvard who call themselves "physicists" in 2097' very probably refers to a hypothesis, but it does not formulate one; for it tells us nothing about the basic principles of the hypothesis (if any) referred to, which hypothesis therefore fails to count as a relevant rival to any current hypothesis.

One especially important consequence of (RR) is that the sheer negation of a hypothesis, unsupplemented by any other claims, does *not* count as a relevant rival to the hypothesis, since the unsupplemented negation of a hypothesis fails to meet condition (a): it cannot sensibly be intended to achieve the theoretical goals of the hypothesis. Simply denying the existence of electrons, for instance, goes no way toward accounting for the phenomena electrons are introduced to explain. Of course, the negation of a hypothesis can certainly be *part* of—one conjunct of—a relevant rival to the hypothesis; and I suspect that this is typically so when it appears that a hypothesis has a pure negation as a rival. So, for example, the denial of electrons' existence conjoined with appropriate phenomenological generalizations can perfectly well be a relevant rival to electron theory. Similarly, while (RR) implies that atheism unadorned is not a relevant rival to theism, it can allow that atheism plus the findings of contemporary science is a relevant rival to theism.

According to (RR), the relevant rivals to a hypothesis will be (1) certain predecessors in the history of the branch of science to which the hypothesis belongs, (2) certain current, actually formulated hypotheses in the branch of science to which the hypothesis belongs, and (3) certain current, actually formulated hypotheses—call them *crackpot hypotheses*—which in some sociological sense do not belong to the branch of science to which the hypothesis belongs (though they may once have done so). So, for instance, to take the SR attitude toward the hypothesis of evolution by gradualist natural selection is (a) to regard the hypothesis as true or false in virtue of the way the mind-independent world is, and (b) to regard it as likelier to be true than (at least) Lamarckianism, the punctuated equilibrium model, and creationism.

In the light of the stipulative definitions (SR) and (RR), it is now tolerably clear at least what (P2) and (P3) are claiming. But are they true? Take (P3) first. Since, according to clause (2) of (SR), taking the SR attitude toward a hypothesis requires only regarding it as more likely to be true than its relevant rivals,[9] and since the relevant rivals to a hypothesis do not include the sheer negation of that hypothesis, it is possible to take the SR attitude toward a hypothesis without regarding it as likely, still less very likely, to be true: a hypothesis might be unlikely, and yet still more likely than its relevant rivals.[10] What about (P2)? Is it true, as (P2) asserts, that the SR attitude, as stipulatively defined, can be identified with the attitude that those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses? I shall now argue that it can.

---

[9] Notice that although regarding a hypothesis as more likely than its relevant rivals may not require being able to estimate *how* likely the hypotheses in question are, it is quite consistent with being able to do so; nothing here assumes or implies that "the testing of theories yields only a comparative warrant"—see Peter Lipton, "Is the Best Good Enough?" in David Papineau, ed., *The Philosophy of Science* (New York: Oxford, 1996), p. 93.

What is the deepest intuitive commitment of those who would call themselves scientific realists and antirelativists? It consists, I suggest, in the respectful way in which they regard (certain) current scientific hypotheses. By and large, they regard these current scientific hypotheses as:

(1) true or false in virtue of the way the mind-independent world is
(2) objectively superior, in some truth-connected sense, to earlier hypotheses in the field, so that science has, in this sense, progressed
(3) objectively superior, in the same sense, to current rival scientific hypotheses
(4) objectively superior, in the same sense again, to current rival hypotheses advocated by people outside the scientific establishment
(5) such that whether the regard for a hypothesis embodied in (1)-(4) is appropriate is generally independent of whether or not the hypothesis postulates entities and properties that cannot be observed

Someone with these attitudes to the best of current scientific hypotheses will therefore find repugnant each of the following three ideas: (a) the (allegedly) neo-Kantian idea that the postulates of a scientific hypothesis are somehow conjured into existence by the widespread acceptance of the hypothesis; (b) the epistemologically egalitarian idea that all hypotheses, past or present, scientific or fringe, are more or less on a cognitive par with one another (even if their *political* influence is unequal); and (c) the empiricist idea that the distinction between observable and unobservable marks a distinction of great epistemological significance.

The argument for (P2)—that is, for identifying the stipulatively-defined SR attitude with the attitude that those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses—has two premises. The first premise is that the attitude that those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses is adequately characterized by (1)-(5). (One who denies this premise has only to specify something that has been left out or wrongly included.) The second premise is that the SR attitude, as stipulatively defined, should be identified with the attitude characterized by (1)-(5). Five observations together provide evidence for this identification. First, clause (1) of (SR) says exactly what (1) above says. Second, to take the SR attitude toward a hypothesis is surely *one* way of taking a truth-connected attitude (as mentioned in (2)-(4)) toward it. It is not, of course, to have an all-or-nothing belief that the hypothesis is true; nor is it even to believe that the hypothesis is closer to the truth than—enjoys greater verisimilitude than—other false hypotheses. But it is, in part, to assign to the hypothesis a higher probability of being *true* than is assigned to its relevant rivals. Third, to take the SR attitude toward a hypothesis, and in particular to assign it a higher probability than any of its relevant rivals, is surely *one* way of regarding it as objectively superior (as mentioned in (2)-(4)) to certain rivals, at least on the assumption that assignments of probability are answerable to objective constraints, such as would be supplied by a reliabilist theory of confirmation (or by a Bayesian theory, appropriately supplemented by some account of objective constraints on assigning prior probabilities). Fourth, the relevant rivals that the SR attitude concerns include just the sort of rivals that (2)-(4) concern. Finally, nothing in the SR attitude rules it out that one could perfectly well take the SR attitude toward a hypothesis postulating unobservables.

---

[10] If the relevant rivals of a hypothesis did include the sheer negation of the hypothesis, then taking the SR attitude toward a hypothesis—regarding it as likelier than all its relevant rivals—would entail regarding it as likelier than its negation, and hence regarding it as likelier than not, and hence as likely.

(P2)'s identification of the SR attitude with the attitude that those with broadly scientific realist and antirelativist sympathies take toward the best of current hypotheses evidently rests upon the idea that nothing compels a scientific realist/antirelativist to assign a high probability to the theories she picks out as best. Yet it is hard, I admit, to dislodge the intuition that *something* requires the assignment of a high probability. But what? Not the need to reject the egalitarian suggestion that all theories, past or present, scientific or fringe, are really on a cognitive par with one another, since that suggestion can be rejected merely by adopting an attitude that, like the SR attitude, only assigns a higher, not a high, probability to favored hypotheses. Here is a better suggestion: "A scientific realist/antirelativist must *believe* the theories she picks out as best, and since to believe a theory just is to assign it a high probability, the high probability requirement swiftly follows." Indeed it does, but neither premise is very plausible. To begin with, it is not at all clear why a scientific realist/antirelativist must *believe* the theories she picks out as best; certainly, belief is not required to explain practical reliance on the theories; for a sufficient basis for action can be the assignment of only a low probability, as when I carry a spare tire, even though I certainly do not believe that I will have a flat. Whereas it might plausibly be claimed that scientific realists/antirelativists must in *some* sense accept the theories they favor, argument is needed to show that this attitude of acceptance amounts to anything more than the SR attitude. More seriously, the identification of belief with the assignment of a high probability runs into the problem of the Lottery Paradox, in which, given the identification, an apparently quite rational person must be regarded as holding contradictory beliefs: a person assigns a high probability to each proposition saying, of one lottery ticket, that it will lose, and also to the proposition that some ticket will not lose; but if belief just is the assignment of high probability, then the person must *believe* of each ticket that it will lose, and also that some ticket will not lose—which commits her to a contradiction.[11] Nor is it true that the assignment of a high probability is even a logically necessary condition of belief: surely, it is logically possible for someone who believes ten (probabilistically independent) propositions, to each of which he assigns a probability of 0.9, to believe also the conjunction of those propositions, even though the probability of the conjunction is low, being the product of the probabilities of each conjunct.[12]

Here is a second suggestion for supporting a high-probability requirement: "Even if believing a theory does not logically require assigning it a high probability, *rationally* believing it surely does. So let us assume some analysis of belief which makes belief quite independent of probability assignments.[13] Then, since a scientific realist/antirelativist must believe the theories she picks out as best, and since *rationally* believing a theory requires assigning it a high probability, a scientific realist/antirelativist must on pain of irrationality assign a high probability to her favored theories." But this second argument for a high-probability requirement is also inconclusive. It retains the nonobvious assumption of its predecessor that a scientific realist/antirelativist must believe the theories she picks out as best. But the premise that rational belief requires assigning a high probability seems clearly wrong; for surely it is always rational to believe the immediate logical consequences of what one rationally believes. But then the person who rationally believes ten (probabilistically independent) propositions, to each of which he assigns a probability of 0.9, is rational in believing their conjunction, even though, as already noted, this conjunction has a low probability.[14] Let me also note that there are theories of rational acceptance which leave open the possibility of accepting hypotheses that are improbable, and they could presumably be modified to cover rational belief, non-probabilistically-construed.[15] Neither Mark Kaplan's nor Patrick Maher's theory is at all reliabilist in spirit, but a reliabilist theory which also yields the same result is easy to imagine in vague outline: if one's cognitive goal is not just truths, but truths that, say, provide a basis for prediction and explanation, then a belief-forming method could be rational to adopt, and its products count as rational, even though unlikely to be true, just so long as they made up for their improbability, as it were, by constituting a superior basis for prediction and explanation. Unformulated theories, or the sheer negations of formulated theories, might be likelier to be true, but cannot provide such a basis.[16]

---

[11] For elaboration and defense, see Mark Kaplan, *Decision Theory as Philosophy* (New York: Cambridge, 1996), pp. 93-98; see also Patrick Maher, *Betting on Theories* (New York: Cambridge, 1993), pp. 134-35.

[12] For the argument as applied to acceptance, see Maher, pp. 137-39 and 152-55.

[13] For such analyses of a belief-like notion of acceptance, see, for example, Kaplan, ch. 4, and Maher, ch. 6.

In the absence of any further reason for insisting on a high-probability requirement, (P2)'s identification of the SR attitude with the attitude that those who have broadly scientific realist and antirelativist intuitions take toward what they regard as the best of current scientific hypotheses can stand.[17] It may be possible to go on the offensive here, however. Resisting a high-probability requirement offers scientific realists/antirelativists an attractive reply to the so-called pessimistic induction which claims that, since most past theories have turned out to be false, current theories are probably false, too.[18] Moreover, it is arguable that insisting on a high-probability requirement would be unreasonable, so that the SR attitude toward hypotheses (or something like it) is all we can decently hope for. Suppose a scientific realist/antirelativist insists that we should regard our best current theories as more likely to be true than false. That implies regarding them as more likely to be true than the disjunction of their rivals, and hence more likely than each disjunct taken individually. But some of those disjuncts are unborn hypotheses, that is, hypotheses that have not yet been formulated and perhaps never will be.[19] How could our current evidence make it reasonable to regard a current hypothesis as likelier than an unformulated hypothesis? So we may be going beyond any attitude it could possibly be reasonable to take if we claim that a current hypothesis is likelier to be true than not.

---

[14] Again, see Maher, pp. 137-39 and 152-55.

[15] See Kaplan, ch. 4, and Maher, ch. 6.3. This is a good point at which to respond to the complaint that my answer to the Hempelian dilemma offers no solace to the physicalist who *believes* physicalism (rather than just taking the SR attitude toward it). In one way, of course, the complaint is misconceived: the thrust of the entire paper is that a physicalist does not have to *believe* physicalism. A more substantial reaction is to note that, even if it is insisted that a physicalist is one who believes physicalism, Hempel's dilemma can be avoided so long as belief in physicalism can be rational despite physicalism's improbability, which would be true if some general thesis held that belief in an improbable hypothesis can still be rational.

[16] Perhaps scientific realist/antirelativist intuitions include the idea that the best of current scientific hypotheses are things we *know*. Perhaps; but it is far from clear that one knows that *p* only if it is likelier to be true than false that *p*. These are accounts of knowledge (for example, reliabilist ones) that do not imply that this is so, and a staunch defender of scientific knowledge might favor such an account—indeed, might be expected to do so.

[17] A different kind of objection to (P2) claims that, if a hypothesis were very, very bad but still better than its relevant rivals, one could take the SR attitude toward it without also favoring it in the manner distinctive of a scientific realist/antirelativist. But, naturally, I doubt that this could happen, and I know of no example. I conjecture that any putative example would either tacitly assume that favoring a hypothesis in the distinctive scientific realist/antirelativist manner requires believing it (which I have denied), or would be a case in which many relevant rivals no less likely could easily be formulated (even though ordinarily no one would bother to).

III

My response to Hempel's dilemma is now complete. Let physicalism be formulated in terms of current physics. Then, given that a physicalist is simply someone who takes the SR attitude toward physicalism, the mere fact that the history of physical theorizing makes physicalism unlikely to be true provides no reason by itself to abandon being a physicalist; one can remain a physicalist, just so long as physicalism, though unlikely, is still more likely than its relevant rivals. This is not simply special pleading on behalf of physicalism; for the SR attitude, as I have argued, is none other than the attitude scientific realists/antirelativists take toward their favored hypotheses in particular branches of science.

Of course, the history of physical theorizing still constitutes evidence against physicalism, in the sense of lowering the likelihood that it is true, and presumably it does not in the same way constitute evidence against physicalism's relevant rivals. So it still threatens physicalism. But any evidence against physicalism must obviously be weighed against evidence for it, and the balance of probabilities may yet leave physicalism likelier than its relevant rivals. Naturally, physicalists insist that there *is* strong evidence for physicalism, evidence capable in principle of giving the probability of physicalism the necessary boost. Such counteracting evidence, moreover, does not have to raise physicalism's probability to 0.5 or higher; given my proposed account of what it is to be a physicalist, it need only lift it above that of its closest relevant rival, an easier requirement to fulfill.

---

[18] As noted earlier, I discovered, in Maher, p. 137.

[19] See Lawrence Sklar, "Do Unborn Hypotheses Have Rights?" *Pacific Philosophical Quarterly*, LXII (1979): 17-29.

In fact, and contrary to the presumption of the last paragraph, the history of physical theorizing may be as damaging to physicalism's relevant rivals as it is to physicalism. Relevant rivals to physicalism are actually formulated hypotheses that are sensibly intended to address the problem that it is the central theoretical goal of physicalism to address: giving an account of the relations among the ontologies that the many sciences (including folk psychology and folk physics) respectively postulate, in light of such cross-scientific regularities as have been discovered empirically.[30] An example of one sort of crude cross-scientific regularity might be that nothing is ever in a mental state unless it is in some simultaneous brain state; but the sciences (for example, the neurosciences) present more refined regularities. To identify physicalism's relevant rivals, it is helpful to view physicalism as the conjunctive thesis that:

(1) There is some science, *S*, distinct from the totality of all the sciences, such that every entity (property) is either itself mentioned as such in the laws and theories of *S* or is ultimately constituted (realized) by entities (properties) mentioned as such in the laws and theories of *S*.

and that:

(2) *S* is current physics.[31]

---

[30] See my *op. cit.*, pp. 222-24.

[31] Understanding physicalism in this way, as the conjunction of (1) and (2), has the possibly unsettling consequence that, say, Thomas Hobbes was not a physicalist, since he had no notion of current physics. But, surely, Hobbesian materialism nevertheless has much in common with physicalism. The most important commonality is the belief that some science distinct from the bare conjunction of the many sciences is basic in some metaphysical sense. But there is also the fact that Hobbes picked as the basic science something that turned out to be, sociologically and historically speaking, the ancestor of current physics. Similarly, if tomorrow new evidence forces a revision of physics, it will be possible to formulate another relevant rival to physicalism in category (A), by letting tomorrow's (superior) physics take the place of current physics in (2); and if the evidence for taking tomorrow's physics to be basic is as good as the evidence for taking current physics to be basic, which is highly likely, the resulting doctrine will be superior. But it will not, strictly speaking, be the doctrine I have called physicalism, though it will, however, have much in common with it. Specifically, it will implement the generic idea that there is a basic science distinct from the bare conjunction of the many sciences, and it will do so by appeal to a science that is an immediate descendent of current physics. So if we are looking for something to serve as "the spirit of physicalism" which transcends particular formulations, then commitment to (1) might be the heart of a good candidate, though the spirit of physicalism would then be cognitive and not attitudinal (see van Fraassen, pp. 169-70).

Relevant rivals to physicalism therefore fall into two categories: (A) those which endorse (1) without endorsing (2), by agreeing that there is a basic science to which all the other sciences stand in some special relation, while proposing that this basic science is something other than physics; and (B) those which deny (1), maintaining that no science is basic, since all sciences are on an ontological par, linked to one another merely by various fundamental laws.[22] Do relevant rivals of either sort gain any advantage over physicalism from evidence constituted solely by the history of physical theorizing?

Take rivals in category (A). Suppose one such rival asserts that there is a basic science, but that it is, say, biology. Such a view is evidentially quite untouched by the history of physical theorizing. On the other hand, the track record of biological theorizing is arguably no better than that of physical theorizing, which evens things out, and obviously there is other evidence decisive against taking biology to be the basic science. But the best-known relevant rival in category (A) is traditional dualism, which I interpret as the view that, to put it very crudely, physicalism is true of everything except the mind: there is a basic science, but it is the *conjunction* of physics and folk psychology (presumably linked to one another by fundamental psychophysical laws). The impact of the history of physical theorizing on traditional dualism, relative to that on physicalism, is trickier to assess. To the extent that the history of physical theorizing makes it likely that current physics is *false*, there is exactly the same evidence against traditional dualism as there is against physicalism, given my interpretation of traditional dualism as the view that the conjunction of physics and folk psychology constitutes the basic science. It might appear, though, that historical evidence that current physics is *incomplete* would leave traditional dualism unharmed, since traditional dualism, unlike physicalism, is not committed to regarding current physics as complete. But, in fact, what the history of physical theorizing makes likely is that current physics has left out something like a new kind of particle, with mass and charge, and traditional dualism *is* committed to regarding current physics as complete in respect of that sort of thing.

What about relevant rivals in category (B)? The egalitarian and pluralist view that there is no basic science, the view, in effect, that cross-scientific regularities should be treated as fundamental laws, appears initially to be committed neither to the truth nor to the completeness of current physics; so on evidence constituted solely by the history of physical theorizing, physicalism appears less likely than antiphysicalist pluralism. Actually, though, on pain of simply saying nothing at all about interscientific relations, antiphysicalist pluralism is committed to law statements connecting events as characterized by current physics to events as characterized by each of the special sciences; and if current physics is probably false, then so, surely, are those law statements. Surprisingly, then, antiphysicalist pluralism may derive no advantage over physicalism from evidence constituted solely by the history of physical theorizing.

---

[22] See Crane and Mellor.

Finally, let me give the briefest outline—no more than a hint—of the evidence *for* physicalism that I mentioned a few paragraphs back. The evidence for physicalism, as I understand it, consists in empirically discovered regularities of various sorts, the best explanation of which is physicalism. One important sort of regularity is that which links physical conditions with simultaneous nonphysical (for example, mental, biological, geological) conditions; so, for instance, there are physical conditions of a person's brain and environment which are regularly accompanied by a certain simultaneous mental condition.[23] Now, physicalism can explain a regularity of this kind by hypothesizing that the mental condition is identical with a functional condition, and that this functional condition is physically realized, and physically realized, in particular, by the physical condition of the person's brain and environment empirically found to be sufficient for it. On this hypothesis, we would expect to observe that whenever the physical condition obtains, so, too, does the mental condition: the physical condition suffices for the playing of a certain functional role, and that something plays that role suffices for the obtaining of the mental condition, so that the physical condition suffices for the mental condition. (Evidently, the physical condition must be *capable* of playing the relevant role, and whether it is must be determined empirically, or else the explanation must be rejected.) But if physicalism is not just an explanation but also the best explanation of such a regularity, then its power to explain the regularity is some evidence for it.[24]

---

[23] There are other empirically-discovered regularities that can arguably be explained by physicalism but not otherwise; for instance, I believe the argument from overdetermination for physicalism (see, for example, my *op. cit.*, pp. 226-31) points us toward some. Yet other regularities potentially providing evidence for physicalism are the laws of the special sciences; see my "Two Cheers for Reductionism; Or, The Dim Prospects For Non-Reductive Materialism," *Philosophy Of Science*, LXII (1995): 370-88.

[24] For elaboration, see my "Being a Physicalist: How and (More Importantly) Why," pp. 231-35.

Further light can be shed on how regularities of this kind provide evidence for physicalism by considering the egalitarian and pluralist rival to physicalism which maintains that there is no basic science. Since this rival asserts the holding of, and hence (trivially) entails, the very regularities that I claim provide evidence for physicalism, it might be objected that these regularities therefore fail to provide evidence that favors physicalism over its rival; the evidence is neutral between physicalism and its rival. But there are obvious replies to this objection. The first is that the rival is still inferior to physicalism in respect of *explanatory power*. Although the egalitarian rival indeed entails the regularities that physicalism can explain, it entails them quite trivially, by explicitly asserting that they hold; so it does not *explain* them at all, but instead treats them as brute, fundamental laws, on a par with the fundamental laws of physics. A second reply is that, both because the egalitarian rival treats these regularities as brute laws to be accepted *in addition to* the fundamental laws of physics, and because it postulates so many properties neither identical with nor even realized by physical properties, it is less *economical* than physicalism. Certainly, antiphysicalists are aware of these replies, but they tend to disparage them. So here is a reason to take them seriously. On my account of the evidence for physicalism, the empirical case for physicalism is strongly analogous to that for any physical theory aiming to give an account of what underlies some (relatively) observable regularity (for example, a physical theory aiming to explain why water boils at 100° C). Now, a physical theory will always have an empirically equivalent rival, namely, the view that the observable regularities it aims to explain should be treated as brute laws that simply have no explanation. But many of us suppose that the theory is nevertheless to be preferred, on the grounds of some superempirical virtue it enjoys, like explanatory power or economy; and whereas the philosophical understanding and justification of appeals to such virtues is obviously contested, that we make such appeals, and take ourselves to be reasonable in doing so, seems hard to deny.[25] But if these appeals are legitimate when used to break the tie between a physical theory and its empirically equivalent phenomenological rival, then why not also in the case of physicalism and its pluralist and egalitarian rival? For the case is simply another instance of a theory and its empirically equivalent rival.[28]

---

[25] I suspect that hostility to taking the superempirical virtues of a theory as counting in favor of its truth stems from the unstated assumption that methodological principles must be a priori in character; for it must be conceded that it is no necessary truth that the world exhibits the sort of economy and connectedness of structure that would make the explanatory power and economy of a theory count in favor of its truth. But from the perspective of an a posteriori, reliabilist epistemology, surely the matter looks quite different.

At first sight, Hempel's dilemma seems to guarantee a swift and decisive victory to the enemies of physicalism. But closer examination reveals that it is merely one offensive in a campaign whose outcome is at best undecided.

ANDREW MELNYK

University of Missouri/Columbia

---

[28] See Smart, "Sensations and Brain Processes," *Philosophical Review*, LXVIII (1959): 141-56, especially pp. 155-56.

BOOK REVIEWS

The Unity of the Self. STEPHEN L. WHITE. Cambridge: Massachusetts Institute of Technology Press, 1991. 424 p. Cloth $42.50.

This volume of essays covers a wide range of topics: content, consciousness, personal identity, self-deception, moral responsibility, moral rightness, and justice. Stephen L. White defends a functionalist view of content and consciousness; argues that personal identity depends on social norms and practices; proposes a homuncular theory of self-deception; presents a novel account of moral responsibility; defends a rights-based approach to morality; and argues that John Rawls is committed to a strategy—namely, "internalizing the veil of ignorance"—that cannot succeed in justifying would-be principles of justice. In the introduction to the volume, White attempts to show that his views on all of these topics bear on issues about the unity of the self. While the attempt is sometimes strained, the breadth of White's philosophical vision is impressive. Moreover, the individual essays exhibit depth and insight. Fundamental issues are addressed with ingenuity, analytical skill, and honesty. The result is a loosely connected volume of high quality essays.

Approximately half of the essays are reprinted. Two of the reprinted essays, "Partial Character and the Language of Thought" (chapter 1) and "Curse of the Qualia" (chapter 3), have had a significant impact in the philosophy of mind. They provide the theoretical foundation—a functionalist view of the mind—for many of the essays in the volume. Because of their impact and foundational role, I shall focus on the main ideas in them and in two supporting essays (namely, "Narrow Content and Narrow Interpretation" (chapter 2) and "Transcendentalism and Its Discontents" (chapter 4)).

CONTENT

White takes the Hilary Putnam-Tyler Burge[1] twin-earth thought experiments to support the thesis that the contents we ordinarily ascribe to beliefs and desires using that-clauses depend on environmental factors, and thus fail to supervene on current, intrinsic physical properties of the subject. He maintains, however, that intentional mode properties such as being a belief and being a desire are determined by functional roles that supervene on current, intrinsic

---

[1] Putnam, "The Meaning of 'Meaning'," reprinted in Putnam, *Mind, Language, and Reality, Volume 2: Philosophical Papers* (New York: Cambridge, 1975), pp. 215-71; and Burge, "Individualism and the Mental," in P. French, et alia, eds., *Midwest Studies in Philosophy, Volume IV* (Minneapolis: Minnesota UP, 1979), pp. 73-121.

0022-362X/97/9412/638–44

physical properties.[2] Moreover, he claims that in addition to having "wide contents," beliefs and desires also have "narrow contents," contents that supervene on functional roles.

White proposes two notions of narrow content. His first and most influential notion is that of partial character (see chapter 1).[3] A partial character is a function from possible contexts of acquisition of a functional state to characters, which are functions from contexts of occurrence to intensions. Partial characters thus represent what characters a functional state would have in various contexts of acquisition. According to White, the intension of a belief or desire (types of functional states) is a function from possible worlds to truth or satisfaction values; the intension of a concept (also a kind of functional state) is a function from possible worlds to extensions. The context of acquisition of a belief or desire state is a function of the contexts of acquisition of its constituent concepts (these contexts can include causal relations and even social factors) and its constituent structure.[4] The partial character of a concept is, then, a function from contexts of acquisition to characters that are functions from contexts of occurrence to functions from possible worlds to extensions. The partial character of a belief or desire is a function from contexts of acquisition to characters that are functions from contexts of occurrence to functions from possible worlds to truth or satisfaction values.

The partial character of a belief or desire is not a kind of propositional content, and thus not something in virtue of which beliefs figure in logical or epistemic relations, or something in virtue of which beliefs and desires figure in practical reasoning or decision making (54). For these reasons, and because of other sources of skepticism about the theoretical interest of the notion of partial character to either folk or scientific psychology, White proposes a second notion of narrow content (see chapter 2). The second notion is a notion of propositional content, where propositional content is a set of possible worlds. Appealing to the notion of the objective optimality of an action, White attempts to show how states can be assigned sets of possible worlds in virtue of their functional roles—he calls such sets, *sets of notional worlds*.[5] While, curiously, White never mentions this, if his attempt succeeds, then any belief or desire will be associated with two propositions (two sets of possible worlds): it will be associated with one (its narrow propositional content) in virtue of its functional role and the other (its wide propositional content) in virtue of its functional role and actual context of occurrence (which will include environmental factors). A belief or desire state would thus have both a wide and a narrow intension.[6]

---

[2] Hereafter by 'functional roles' I shall always mean functional roles which are "narrow" in that they supervene on current, intrinsic properties of the subject, and which are specifiable in physical and/or topic-neutral terms.

[3] The notion is so-called because it was developed by analogy with David Kaplan's notion of the characters of words and sentences presented in his "On the Logic of Demonstratives," *Journal of Philosophical Logic*, VIII (1978): 81-98. Several nonequivalent notions of narrow content can be found in the literature on intentionality. White's notion of partial character has come to be one of the leading ones. (Jerry Fodor embraced White's notion of partial character in *Psychosemantics* (Cambridge: MIT, 1987), though he nowadays eschews any notion of narrow content.)

[4] White is sympathetic to the view that concepts are words in a language of thought (47).

White boldly contends that narrow propositional contents are of theoretical interest because belief and desire states satisfy certain rationality constraints in virtue of such contents (67). The rationality constraints will be expressed "in a substantive theory of rationality" that might be "developed along the lines of Baysian decision theory" (67). The theory will entail that "subjects maximize their expected utility subject to computational limitations and to relatively local lapses into irrationality" (67). White contends that "the propositional attitude ascriptions that make subjects maximizers of expected utility are narrow content ascriptions" (69). The justification for positing such narrow contents lies, he maintains, "in our interest in wide content coupled with the role that rationality constraints, and thus narrow content play on its production" (72).

I think that White's account of how states can be assigned sets of possible worlds in virtue of their functional roles is unsuccessful. But since the account is complicated and details are missing, rather than pursue this point, I shall instead say why even if the account were successful, White fails to show that beliefs and desires satisfy rationality constraints in virtue of narrow propositional contents. Some of the considerations that seem to indicate that beliefs and desires do not satisfy rationality constraints in virtue of their characters or wide intensions seem to indicate, as well, that they do not satisfy them in virtue of narrow propositional contents either. The that-clause ascriptions under which beliefs and desires satisfy rationality constraints are hyperintensional: terms with the same characters and/or wide intensions can fail to be substitutable *salva veritate* in them. Wide intensions and characters fail to individuate beliefs and desires finely enough to capture their roles in decision-theoretic relations. White offers no reason to believe that terms that are co-extensive in every world in the set of notional worlds which is the narrow propositional content of a belief or desire are substitutable *salva veritate* in such that-clause ascriptions; and he offer no pragmatic account of how such failures of substitution can be explained away. He thus fails to make a case that individuating beliefs and desires by narrow propositional contents captures their roles in decision-theoretic relations.

---

[5] He so calls them because this second notion of narrow content is inspired by Daniel Dennett's idea of notional worlds in his "Beyond Belief," in Andrew Woodfield, ed., *Thought and Object* (New York: Oxford, 1982); White's notion is interestingly different from Dennett's, however.

[6] In *The Conscious Mind* (New York: Oxford, 1996), David Chalmers claims that beliefs and desires have two kinds of intensions: primary intensions (sets of centered possible worlds) and secondary intensions (sets of possible worlds). Moreover, he maintains that primary intensions are narrow, and that secondary intensions are wide. While Chalmers's notion of primary intensions and White's notion of narrow propositional contents are interestingly related in other ways as well, Chalmers's notion differs from White's in ways I lack the space to discuss here.

Unfortunately, White says nothing at all about the relationship between this notion of narrow propositional content and the notion of partial character. The reader is left wondering, for instance, whether he mistakenly thinks there is a one-one correspondence between the partial characters of beliefs and desires and their narrow propositional contents. In any case, call the properties we ascribe to belief and desire states using the hyperintensional that-clause ascriptions under which they satisfy rationality constraints, *hyperintensional contents*. White could ensure that the functional role contribution of a concept to hyperintensional contents supervenes on its partial character by stipulating that the context of acquisition for the concept include whatever functional-role aspects (additional to those which make the state a concept) contribute to hyperintensional contents. *If*, then, the functional-role contribution of concepts exhausts their internal (that is, nonenvironmental) contribution to hyperintensional contents, the partial characters of concepts would at least serve the following theoretical purpose: they would represent the internal contribution of concepts to the properties in virtue of which beliefs and desires satisfy rationality constraints. Moreover, *if* properties such as having a certain propositional content supervene on the partial characters of beliefs and desires, then these narrow propositional contents will supervene on the internal contributions of beliefs and desires to the satisfaction of rationality constraints.

QUALIA

The *apparent* possibility of absent and inverted qualia poses a prima facie problem for orthodox functionalism, a doctrine that implies that it is a priori true that qualitative properties of states supervene on functional properties (76). In defense of orthodox functionalism, White makes a two-part case that the intuition that absent and inverted qualia are possible should be abandoned. Each part of the defense, I believe, fails.

First, he argues as follows that the rejection of orthodox functionalism implies that there are irreducible mental properties.

> *Premise one*: "if two expressions refer to the same object [or state] and this fact cannot be established a priori, they do so in virtue of different routes to the referent provided by different modes of presentation of that referent" (92).
> *Premise two*: "modes of presentation are properties" (92).
> *Premise three*: Mental descriptions are not a priori coreferential with any physical descriptions.
> *Conclusion*: "if [contra orthodox functionalism] there are no topic-neutral expressions that are at least coreferential a priori with such mentalistic descriptions as 'Smith's pain at t', then these mentalistic descriptions refer in virtue of a property distinct from that in virtue of which any physicalist or topic-neutral expression refers" (94).

The premises fail to support this conclusion. At best they support the claim that any properties associated with such descriptions as 'Smith's pain at *t*' as modes of presentation are distinct from any property that is associated with a physical and/or topic-neutral description as a mode of presentation. To derive the conclusion that "mentalistic descriptions refer in virtue of a property distinct from that in virtue of which any physicalist or topic-neutral expression refers," White must assume that the mode of presentation associated with a description is invariably the property in virtue of which something satisfies the description; and that assumption is false.

In the first-person case, the mode of presentation associated with 'pain' and the concept this term expresses is pain itself: pain states are self-presenting states. The modes of presentation associated with a physical description such as 'brain state *X*' and the concept it expresses are, in contrast, not the property of being brain state *X*, but rather properties that we take to indicate the presence of brain state *X*, properties such as having such-and-such an effect on an electroencephalograph. The fact that the concept of pain and the concept of brain state *X* have different associated modes of presentation does not imply that pain is not brain state *X*.[7]

---

[7] Christopher Hill offers a similar response to White's argument in *Sensations* (New York: Cambridge, 1991), pp. 98-101.

White's second line of argument (108) is that the basic intuitions that underlie the intuition that absent and inverted qualia are possible lead, upon reflection, to the unacceptable doctrine of transcendentalism—the doctrine that qualitative properties are subjective states that fail to supervene on any objective properties whatsoever (106)—and thus should be rejected.[8] Absent qualia, for instance, seem possible because for any type of functional state, we can both imagine and conceive of an individual in that functional state failing to be in any qualitative state whatsoever. But given the subjective character of qualitative states, for any objective state whatsoever, functional or physical or otherwise, we can both imagine and conceive of an individual in that objective state failing to be in any qualitative state at all.

These claims about imaginability and conceivability seem plausible. Rather than support orthodox functionalism, however, they support the claim that it is not a priori true that qualitative states supervene on objective states, and thus support the denial of orthodox functionalism.

Acknowledging that it is not a priori true that qualitative states supervene on objective states does not commit one to transcendentalism. For it remains an open question whether qualitative properties are identical with physical or functional properties.[9] The concept of water is not a priori equivalent to the concept of H_2O, yet it is impossible for something to be water and not be H_2O. Of course, Saul Kripke[10] noted an important disanalogy between the concept of pain and the concept of water. Something can look, smell, taste, and feel like water yet not fall under the concept of water; whereas whatever feels like pain falls under the concept of pain. So, to be sure, we cannot explain away the appearance of contingency of 'Pain = brain state *X*' (or the like) in the way that Kripke explains away the appearance of contingency of the necessary truth that water = H_2O. Still, there may be some other way of explaining away the appearance of contingency of 'Pain = brain state *X*'.

The physicalist and functionalist should argue that given the conceptual role differences between qualitative concepts and physical and functional concepts, a priori methods such as imaginability and conceivability are unreliable guides to possibility in the cases in question. We have already noted a crucial difference between a qualitative concept, such as the concept of pain, and relevant physical and topic-neutral ones: while pain itself is the associated mode of presentation of the concept of pain, the associated modes of presentation of relevant physical and functional concepts are indirect—they are properties we take to indicate the presence of the properties in virtue of which the concepts apply, not the properties in virtue of which the concepts apply. Consider, then, imaginability.[11] When one imagines—from the first-person point of view—someone in pain, one sympathetically imagines one's being in pain. When one imagines someone in a brain state or in a functional state, one imagines perceiving the instantiation of properties that are indirect modes of presentation of the state. One could imagine the brain state or functional state in question accompanied by the absence of pain by perceptually imagining an indirect mode of presentation of the brain state or functional state while at the same time sympathetically imagining being free of pain. One could so imagine even if pain is identical with the brain state or functional state in question. Physicalists and functionalists should argue that such imaginative acts, involving as they do the joint exercise of the distinct faculties of perceptual and sympathetic imagination, are unreliable guides to possibility.[12]

---

[8] White argues that transcendentalism is unacceptable because (a) it implies that qualitative mental states are epiphenomena, and (b) it cannot explain the sense in which we have privileged access to our qualitative mental states (chapter 4). (Chalmers embraces transcendentalism, together with these consequences of the doctrine.)

[9] Supervenience is a reflexive relation; so identity guarantees supervenience.

[10] *Naming and Necessity* (Cambridge: Harvard, 1980).

BRIAN P. MCLAUGHLIN

Rutgers University

---

[11] Conceivability would require a different, though related treatment. See the references in footnote 12.

[12] The source of this point is footnote 11 of Thomas Nagel's "What Is It Like to Be a Bat?" *The Philosophical Review*, LXXV (1974): 435-50. For further discussion, see Hill's "Imaginability, Conceivability, Possibility, and the Mind-Body Problem," *Philosophical Studies*, LXXXVII (1997): 61-85; and Hill and my "There Are Fewer Things in Reality than Are Dreamt of in Chalmers's Philosophy," forthcoming in *Philosophy and Phenomenological Research*. See also Brian Loar's "Phenomenal States," *Philosophical Perspectives*, IV (1990): 81-108.

INDEX TO VOLUME XCIV

INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

Abortion

Tomasi, John: Liberalism, Sanctity, and the Prohibition of Abortion. 94 (10): 491-513 '97

Bayesian Theory

Eckhardt, William: A Shooting-room View of Doomsday. 94 (5): 244-59 '97

Belief

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
Roorda, Jonathan: Fallibilism, Ambivalence, and Belief. 94 (3): 126-55 '97

Burke, Michael B.

Olson, Eric T.: Dion's Foot. 94 (5): 260-65 '97

Conditionals

Morreau, Michael: Fainthearted Conditionals. 94 (4): 187-211 '97

Davidson, Donald

Sommers, Fred: Putnam's Born-again Realism. 94 (9): 453-71 '97

Deception

Adler, Jonathan E.: Lying, Deceiving, or Falsely Implicating. 94 (9): 435-52 '97

Decision Theory

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Della Rocca, Michael

Baker, Lynne Rudder: Why Constitution Is Not Identity. 94 (12): 599-622 '97

Doom

Eckhardt, William: A Shooting-room View of Doomsday. 94 (5): 244-59 '97

Dworkin, Ronald

Tomasi, John: Liberalism, Sanctity, and the Prohibition of Abortion. 94 (10): 491-513 '97

Elgin, Catherine Z.

van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Epistemology

Fogelin, Robert J.: Quine's Limited Naturalism. 94 (11): 543-63 '97
Roorda, Jonathan: Fallibilism, Ambivalence, and Belief. 94 (3): 126-55 '97
van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Essentialism

Olson, Eric T.: Dion's Foot. 94 (5): 260-65 '97

Ethics

Adler, Jonathan E.: Lying, Deceiving, or Falsely Implicating. 94 (9): 435-52 '97
Thomson, Judith Jarvis: The Right and the Good. 94 (6): 273-98 '97

Fodor, Jerry A.

Blachowicz, James: Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97

Gibbard, Allan

Baker, Lynne Rudder: Why Constitution Is Not Identity. 94 (12): 599-622 '97

Good

Thomson, Judith Jarvis: The Right and the Good. 94 (6): 273-98 '97

Goodman, Nelson

Blachowicz, James: Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97

Hempel, Carl G.

Melnyk, Andrew: How to Keep the 'Physical' in Physicalism. 94 (12): 622-37 '97

Identity

Baker, Lynne Rudder: Why Constitution Is Not Identity. 94 (12): 599-622 '97
Johnston, Mark: Manifest Kinds. 94 (11): 564-83 '97

Identity, Personal

Olson, Eric T.: Dion's Foot. 94 (5): 260-65 '97

Justice

Ackerman, Bruce: Temporal Horizons of Justice. 94 (6): 299-317 '97

Knowledge, Theory of

Sosa, Ernest: Reflective Knowledge in the Best Circles. 94 (8): 431-34 '97

Kripke, Saul

Baker, Lynne Rudder: Why Constitution Is Not Identity. 94 (12): 599-622 '97
Johnston, Mark: Manifest Kinds. 94 (11): 564-83 '97

Leslie, John

Eckhardt, William: A Shooting-room View of Doomsday. 94 (5): 244-59 '97

Levi, Isaac

Roorda, Jonathan: Fallibilism, Ambivalence, and Belief. 94 (3): 126-55 '97

Lewis, David

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
Blachowicz, James: Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97
Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Liberalism

Ackerman, Bruce: Temporal Horizons of Justice. 94 (6): 299-317 '97
Tomasi, John: Liberalism, Sanctity, and the Prohibition of Abortion. 94 (10): 491-513 '97

Logic

Copeland, B. Jack: Vague Identity and Fuzzy Logic. 94 (10): 514-34 '97
Dipert, Randall R.: The Mathematical Structure of the World: The World as Graph. 94 (7): 329-58 '97
Kyburg, Henry E., Jr.: The Rule of Adjunction and Reasonable Inference. 94 (3): 109-25 '97
Morreau, Michael: Fainthearted Conditionals. 94 (4): 187-211 '97

Logic, Philosophy of

Copeland, B. Jack: Vague Identity and Fuzzy Logic. 94 (10): 514-34 '97
Kyburg, Henry E., Jr.: The Rule of Adjunction and Reasonable Inference. 94 (3): 109-25 '97

Morreau, Michael: Fainthearted Conditionals. 94 (4): 187-211 '97

Lying

Adler, Jonathan E.: Lying, Deceiving, or Falsely Implicating. 94 (9): 435-52 '97

Mathematics, Philosophy of

Azzouni, Jody: Thick Epistemic Access: Distinguishing the Mathematical from the Empirical. 94 (9): 472-84 '97
Fallis, Don: The Epistemic Status of Probabilistic Proof. 94 (4): 165-86 '97
Kagan, Shelly: Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97
Vallentyne, Peter: Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97

Metaphysics

Baker, Lynne Rudder: Why Constitution Is Not Identity. 94 (12): 599-622 '97
Dipert, Randall R.: The Mathematical Structure of the World: The World as Graph. 94 (7): 329-58 '97
Johnston, Mark: Manifest Kinds. 94 (11): 564-83 '97
Olson, Eric T.: Dion's Foot. 94 (5): 260-65 '97

Mind, Philosophy of

Blachowicz, James: Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97
Cummins, Robert: The Lot of the Causal Theory of Mental Content. 94 (10): 535-42 '97
Hardcastle, Valerie Gray: When a Pain Is Not. 94 (8): 381-409 '97

Moore, G. E.

Sosa, Ernest: Reflective Knowledge in the Best Circles. 94 (8): 431-34 '97
Thomson, Judith Jarvis: The Right and the Good. 94 (6): 273-98 '97

Naturalism

Fogelin, Robert J.: Quine's Limited Naturalism. 94 (11): 543-63 '97

New Books

94 (6): 325-28
94 (7): 375-78

New Books: Anthologies

94 (2): 107
94 (7): 378-80
94 (11): 597-98

New Books: Translations

94 (2): 108
94 (5): 272

Notes and News

94 (7): 374

Pain

Hardcastle, Valerie Gray: When a Pain Is Not. 94 (8): 381-409 '97

Philosophy, Political: *See* Political Theory.

Physicalism

Fogelin, Robert J.: Quine's Limited Naturalism. 94 (11): 543-63 '97
Melnyk, Andrew: How to Keep the 'Physical' in Physicalism. 94 (12): 622-37 '97

Political Theory

Ackerman, Bruce: Temporal Horizons of Justice. 94 (6): 299-317 '97

Probability

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97
Fallis, Don: The Epistemic Status of Probabilistic Proof. 94 (4): 165-86 '97
Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Putnam, Hilary

Sommers, Fred: Putnam's Born-again Realism. 94 (9): 453-71 '97 van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Quine, W. V.

Azzouni, Jody: Thick Epistemic Access: Distinguishing the Mathematical from the Empirical. 94 (9): 472-84 '97

Fogelin, Robert J.: Quine's Limited Naturalism. 94 (11): 543-63 '97

Ramsey, Frank

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Realism

Sommers, Fred: Putnam's Born-again Realism. 94 (9): 453-71 '97

van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Representation

Blachowicz, James: Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97

Cummins, Robert: The Lot of the Causal Theory of Mental Content. 94 (10): 535-42 '97

Right

Thomson, Judith Jarvis: The Right and the Good. 94 (6): 273-98 '97

Salmon, Wesley

Beebee, Helen: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Papineau, David: Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Science, Philosophy of

Melnyk, Andrew: How to Keep the 'Physical' in Physicalism. 94 (12): 622-37 '97

van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Truth

Sommers, Fred: Putnam's Born-again Realism. 94 (9): 453-71 '97

van Fraassen, Bas C.: Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Vagueness

Copeland, B. Jack: Vague Identity and Fuzzy Logic. 94 (10): 514-34 '97

Value, Theory of

Kagan, Shelly: Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97

Vallentyne, Peter: Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97

INDEX OF AUTHORS

Including Contributing Authors and Reviewers

Ackerman, Bruce, Temporal Horizons of Justice. 94 (6): 299-317 '97

Adler, Jonathan E., Lying, Deceiving, or Falsely Implicating. 94 (9): 435-52 '97

Amdur, Robert, Review of Kent Greenawalt: *Private Consciences and Public Reasons*. 94 (11): 584-87 '97

Azzouni, Jody, Thick Epistemic Access: Distinguishing the Mathematical from the Empirical. 94 (9): 472-84 '97

Baker, Lynne Rudder, Why Constitution Is Not Identity. 94 (12): 599-622 '97

Beebee, Helen, Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Blachowicz, James, Analog Representation beyond Mental Imagery. 94 (2): 55-84 '97

Copeland, B. Jack, Vague Identity and Fuzzy Logic. 94 (10): 514-34 '97

Cummins, Robert, The Lot of the Causal Theory of Mental Content. 94 (10): 535-42 '97

Dipert, Randall R., The Mathematical Structure of the World: The World as Graph. 94 (7): 329-58 '97

Dummett, Michael, Review Essay of Pasquale Frascolla: *Wittgenstein's Philosophy of Mathematics*; and Christoffer Gefwert: *Wittgenstein on Philosophy and Mathematics*. 94 (7): 359-74 '97

Dworkin, Gerald, Review of Bernard Berofsky: *Liberation from Self: A Theory of Personal Autonomy*. 94 (4): 212-16 '97

Eckhardt, William, A Shooting-room View of Doomsday. 94 (5): 244-59 '97

Eldridge, Richard, Review of Martha C. Nussbaum: *Poetic Justice: The Literary Imagination and Public Life*. 94 (8): 431-34 '97

Fallis, Don, The Epistemic Status of Probabilistic Proof. 94 (4): 165-86 '97

Fogelin, Robert J., Quine's Limited Naturalism. 94 (11): 543-63 '97

Gauthier, David, Review of Quentin Skinner: *Reason and Rhetoric in the Philosophy of Hobbes*. 94 (2): 94-97 '97

Godfrey-Smith, Peter, Review of Elliott Sober: *From a Biological Point of View: Essays in Evolutionary Philosophy*. 94 (3): 160-64 '97

Hardcastle, Valerie Gray, When a Pain Is Not. 94 (8): 381-409 '97

Hardimon, Michael O., Review of Axel Honneth: *The Struggle for Recognition: The Moral Grammar of Social Conflicts*. 94 (1): 46-54 '97

Humphrey, Nicholas, Review of Daniel C. Dennett: *Kinds of Minds: Toward an Understanding of Consciousness*. 94 (2): 97-103 '97

Johnston, Mark, Manifest Kinds. 94 (11): 564-83 '97

Kagan, Shelly, Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97

Kyburg, Henry E., Jr., The Rule of Adjunction and Reasonable Inference. 94 (3): 109-25 '97

Levi, Isaac, Review of Frederic Schick: *Making Choices: A Recasting of Decision Theory*. 94 (11): 588-97 '97

MacIntyre, Alasdair, Review of Charles Larmore: *The Morals of Modernity* and *The Romantic Legacy*. 94 (9): 485-90 '97

McLaughlin, Brian P., Review of Stephen L. White: *The Unity of the Self*. 94 (12): 638-44 '97

Melnyk, Andrew, How to Keep the 'Physical' in Physicalism. 94 (12): 622-37 '97

Morreau, Michael, Fainthearted Conditionals. 94 (4): 187-211 '97

Nussbaum, Martha C., Review Essay of Gregory Vlastos: *Socratic Studies*. 94 (1): 27-45 '97

Olson, Eric T., Dion's Foot. 94 (5): 260-65 '97

Papineau, David, Probability as a Guide to Life. By Helen Beebee and David Papineau. 94 (5): 217-43 '97

Pippin, Robert B., Review of Béatrice Longuenesse: *Kant et le pouvoir de juger: Sensibilité et discursivité dans l'analytique transcendentale de la* Critique de la raison pure. 94 (6): 318-24 '97

Roorda, Jonathan, Fallibilism, Ambivalence, and Belief. 94 (3): 126-55 '97

Sommers, Fred, Putnam's Born-again Realism. 94 (9): 453-71 '97

Sosa, Ernest, Reflective Knowledge in the Best Circles. 94 (8): 431-34 '97

Sturgeon, Nicholas L., Review of Stephen Darwall: *The British Moralists and the Internal 'Ought': 1640-1740.* 94 (5): 266-71 '97

Thomson, Judith Jarvis, The Right and the Good. 94 (6): 273-98 '97

Tomasi, John, Liberalism, Sanctity, and the Prohibition of Abortion. 94 (10): 491-513 '97

Vallentyne, Peter, Infinite Value and Finitely Additive Value Theory. By Peter Vallentyne and Shelly Kagan. 94 (1): 5-26 '97

van Fraassen, Bas C., Elgin on Lewis's Putnam's Paradox. 94 (2): 85-93 '97

Wallace, R. Jay, Review of John Martin Fischer: *The Metaphysics of Free Will: An Essay on Control.* 94 (3): 156-59 '97

Williamson, Timothy, Review of William P. Alston: *A Realist Conception of Truth.* 94 (2): 103-06 '97

INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

Alston, William P.: *A Realist Conception of Truth.*—Timothy Williamson. 94 (2): 103-06 '97

Berofsky, Bernard: *Liberation from Self: A Theory of Personal Autonomy.*—Gerald Dworkin. 94 (4): 212-16 '97

Darwall, Stephen: *The British Moralists and the Internal 'Ought': 1640-1740.*—Nicholas L. Sturgeon. 94 (5): 266-71 '97

Dennett, Daniel C.: *Kinds of Minds: Toward an Understanding of Consciousness.*—Nicholas Humphrey. 94 (2): 97-103 '97

Fischer, John Martin: *The Metaphysics of Free Will: An Essay on Control.*—R. Jay Wallace. 94 (3): 156-59 '97

Frascolla, Pasquale: *Wittgenstein's Philosophy of Mathematics*—Michael Dummett. 94 (7): 359-74 '97

Gefwert, Christoffer: *Wittgenstein on Philosophy and Mathematics.*—Michael Dummett. 94 (7): 359-74 '97

Greenawalt, Kent: *Private Consciences and Public Reasons.*—Robert Amdur. 94 (11): 584-87 '97

Honneth, Axel: *The Struggle for Recognition: The Moral Grammar of Social Conflicts.*—Michael O. Hardimon. 94 (1): 46-54 '97

Larmore, Charles: *The Morals of Modernity* and *The Romantic Legacy.*—Alasdair MacIntyre. 94 (9): 485-90 '97

Longuenesse, Béatrice: *Kant et le pouvoir de juger: Sensibilité et discursivité dans l'analytique transcendantale de la* Critique de la raison pure.—Robert B. Pippin. 94 (6): 318-24 '97

Nussbaum, Martha C.: *Poetic Justice: The Literary Imagination and Public Life*.—Richard Eldridge. 94 (8): 431-34 '97

Schick, Frederic: *Making Choices: A Recasting of Decision Theory*.—Isaac Levi. 94 (11): 588-97 '97

Skinner, Quentin: *Reason and Rhetoric in the Philosophy of Hobbes*.—David Gauthier. 94 (2): 94-97 '97

Sober, Elliott: *From a Biological Point of View: Essays in Evolutionary Philosophy*.—Peter Godfrey-Smith. 94 (3): 160-64 '97

Vlastos, Gregory: *Socratic Studies*.—Martha C. Nussbaum. 94 (1): 27-45 '97

White, Stephen L.: *The Unity of the Self*.—Brian P. McLaughlin. 94 (12): 638-44 '97

UNITED STATES POSTAL SERVICE™

Statement of Ownership, Management, and Circulation

(Required by 39 USC 3685)

1. Publication Title THE JOURNAL OF PHILOSOPHY	2. Publication Number 0 0 2 2 - 3 6 2 X	3. Filing Date 10/4/97
4. Issue Frequency MONTHLY	5. Number of Issues Published Annually 12	6. Annual Subscription Price $35.00 individuals $75.00 institutions

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4)

THE JOURNAL OF PHILOSOPHY
1150 Amsterdam Avenue
Mail Code 4972
709 Philosophy Hall
Columbia University
New York, NY 10027

Contact Person: Pamela C Ward
Telephone: 212/866-4419

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)

SAME

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address)

THE JOURNAL OF PHILOSOPHY
1150 Amsterdam Avenue
Mail Code 4972
709 Philosophy Hall
Columbia University
New York, NY 10027

Editor (Name and complete mailing address)

Arthur Danto
1150 Amsterdam Avenue
Mail Code 4972
709 Philosophy Hall
Columbia University
New York, NY 10027

Managing Editor (Name and complete mailing address)

Michael Kelly
1150 Amsterdam Avenue
Mail Code 4972
709 Philosophy Hall
Columbia University
New York, NY 10027

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name	Complete Mailing Address
THE JOURNAL OF PHILOSOPHY, INC	1150 Amsterdam Avenue
	Mail Code 4972
	709 Philosophy Hall
	Columbia University
	New York, New York 10027

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☐ None

Full Name	Complete Mailing Address

12. Tax Status (For completion by nonprofit organizations authorized to mail at special rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, September 1995 (See Instructions on Reverse)

13. Publication Title THE JOURNAL OF PHILOSOPHY	14. Issue Date for Circulation Data Below DECEMBER 1997	

15. Extent and Nature of Circulation		Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)		4000	3925
b. Paid and/or Requested Circulation	(1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales (Not mailed)		
	(2) Paid or Requested Mail Subscriptions (Include advertiser's proof copies and exchange copies)	3500	3402
c. Total Paid and/or Requested Circulation (Sum of 15b(1) and 15b(2)) ▶		3500	3402
d. Free Distribution by Mail (Samples, complimentary, and other free)		150	150
e. Free Distribution Outside the Mail (Carriers or other means)			
f. Total Free Distribution (Sum of 15d and 15e) ▶		150	150
g. Total Distribution (Sum of 15c and 15f) ▶		3650	3552
h. Copies not Distributed	(1) Office Use, Leftovers, Spoiled	350	373
	(2) Returns from News Agents		
i. Total (Sum of 15g, 15h(1), and 15h(2)) ▶		4000	3925
Percent Paid and/or Requested Circulation (15c / 15g x 100)		96	96

16. Publication of Statement of Ownership
☒ Publication required. Will be printed in the 12/97 issue of this publication.
☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner Michael Kelly, Managing Editor	Date October 6, 1997

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including multiple damages and civil penalties).